Modelling the Range-wide Density Patterns of the *Arthroleptella lightfooti* using Acoustic Monitoring Data

Jenicca Poongavanan

Thesis presented in partial fulfilment of the requirements for the degree

Master of Science in Ecological Statistics

University of Cape Town

February 2018

Supervisor: Assoc Prof Res Altwegg

Co-Supervisors: Dr. Ian Durbach, Dr. John Measey

Abstract

Species distributions are often limited by environmental factors and according to the abundant-centre hypothesis, abundance should be highest where the environment is most favourable for the species. So, do the same environmental factors determine occurrence and abundance patterns inside the range? I examined this question using *Arthroleptella lightfooti*, a species of frog from the family of Pyxicephalidae, endemic to the mountains of the Cape peninsula, South Africa.

I used density estimates obtained from acoustic Spatially Explicit Capture Recapture (aSCR) methods and data from an acoustic survey using an array of 6 microphones to construct the first Peninsula wide population-density surface for this visually cryptic but acoustically active species. The analysis consisted of three stages. The first involved creating two sets of data from the original: one shows whether the species is present or not and the other indicates the density when the species is present. The second stage consisted of fitting a Hurdle Model to the data where the presence data is modelled using logistic regression and the density data is separately modelled using ordinary linear regression. The third stage involved combining the two models to estimate the expected density of the species. Confidence intervals were built using non-parametric bootstrapping.

It was found that covariates explaining variation in occurrence were not the same as those explaining variation in density, suggesting that processes determining occurrence were not always those determining density. Of the environmental conditions examined, although predictive of occurrence, were generally poor predictors of *A. lightfooti* density. Presence of the Lightfoot's moss frog was largely explained by topographic features and availability of water. In contrast, predictions of density were only weakly related to these same environmental factors and in some cases contradicting one another.

The second part of this study produces the first Peninsula wide population density surface of *A. lightfooti*. At the same time, it assesses the ability of using opportunistically collected presence-only records in combination with the higher quality density data to improve the estimation of expected population-density surface of *A. lightfooti*. The presence-only records were constructed into a habitat suitability map using an ensemble of species distribution models. The habitat suitability map was then integrated in the modelling framework as a covariate in order to improve the estimation of expected population-density surface of *A. lightfooti*. However, the habitat suitability covariate resulted as being uninformative.

Acknowledgement

The first persons I would like to extend my gratitude to are my supervisors: Res Altwegg, for the innumerable times I ran in your office with my obscure questions and for your positive outlook on my research. Ian Durbach, thank you for always making time for skype meetings and for always encouraging me and John Measey for providing the data and for introducing me to the world of amphibians. I would also like to thank all the SEEC members for welcoming me in the statistical ecology research group. I am thankful for the ACDI, NRF and UCT for funding this project. In addition I would like to thank the National Geographic grant (9913-16) for funding the fieldwork associated with this project and to Marike Louw who collected the data used in this study.

And then on a personal note, I would like to extend my utmost gratitude to my family. You have all supported me even by being miles away. You have always driven me to achieve my very best and have stood by me when I thought my best was not good enough. I am eternally grateful for all you have done and sacrificed for me. I would also like to extend my gratitude to my friends (more like family), Juhi Hurgobin and Delan Hassoo. You have supported me through the good times, as well as the crazy and really challenging moments. I would not have made it to the very end without my sense of humour in check. Thank you all!

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

1.1 General Introduction

Knowledge of patterns of species distribution and abundance is a central theme of ecology (Gaston, 2003). Particularly, understanding the spatial and temporal pattern of distribution and abundance is crucial to the conservation and management of threatened species (Nielsen et al., 2005). The availability of comprehensive environmental data, robust statistical method and Geographic information system (GIS) tools have boosted the development of Species distribution models (SDMs, Guisan et al., 2017; Franklin, 2010). These methods use the correlation between species occurrence and environmental variables to map and predict the potential geographical range of a species or explain why a species occurs where it does.

Although species distribution models have largely been used to describe species occurrence, species abundance models (Potts and Elith, 2006) are less common, despite the greater information for conservation management (Nielsen et al., 2005). Species Abundance Models (SAMs) are species distribution models built from abundance (count) data. Prediction patterns of abundance are much more difficult to obtain due to the lack of data on spatial variation in abundance within the species distribution, especially in the case of rare and cryptic species (Sagarin et al., 2006). Consequently inferences on species' habitat preferences were originally drawn on simplistic assumptions about species distribution (Nielsen et al., 2005). Moreover, it is known that species distributions are often limited by environmental factors and according to the abundant-centre hypothesis (Sagarin and Gaines, 2002), abundance should be highest where the environment is most favourable for the species (Péron and Altwegg, 2015). But, do the same environmental factors determining occurrence also determine abundance patterns inside the range?

The abundant centre hypothesis is based on the formulation that species abundance distributions have a positive relationship with environmental gradients, whereby at a given point along this gradient (e.g. temperature) conditions are optimal for the species and this is where it attains its highest population density (Sagarin and Gaines, 2002). Moving away from this optimal point, the species will experience less favourable conditions leading to a decline in population. However, several studies (Filz et al., 2013; Nielsen et al., 2005) identified cases where this hypothesis may simply be unsupported. It might be difficult to find a strong abundance-suitability relationship for species with small ranges. As in such cases, environmental variables might vary within a relatively small range of values. There are some alternative patterns to the abundant centre hypothesis. For example, abundances of some species decline consistently from one limit of their range to the other and are therefore said to have a ramped distribution (Baldanzi et al., 2013).

A perfect example of a cryptically coloured and small species is *Arthroleptella lightfooti* (also commonly known as the Lightfoot's moss frog and hereinafter referred to as *A.lightfooti*). It is a species of frog from the family of Pyxicephalidae, endemic to Table Mountain and other mountains of the Cape peninsula, South Africa (Turner and De Villiers, 2007). Adult females can reach up to 22mm in length while the males are smaller and have a short, insect-like chirp advertisement call consisting of three rapid pulses of 0.1 s long at an accentuated frequency of 3.1-3.4 kHz (Channing, 2001; Measey et al., 2017). It inhabits seepages in fynbos areas (Channing, 2004). Calls peak during breeding seasons between April and December (Minter, 2004). Females lay clutches of 5-12 eggs on damp soil near rivers, roadside seepages and heavily vegetated streams (Channing, 2004). This species does not occur in sympatry with any other *Arthroleptella* species and can be reliably distinguished in the field by their advertisement calls (Turner and De Villiers, 2007; Dawood and Channing, 2000).

Moreover, the species is listed as near threatened by the IUCN (2018). Although inhabiting a largely protected area, *A. lightfooti* faces major threats, including the invasion of alien plants and frequent and intense fires (Measey and Tolley, 2011). Previously no conservation actions were prioritised for this species, however, no accurate population density estimates of the species were known. The following section gives a brief overview of sampling methods and their extensions which led to the availability of a first estimation of the *A. lightfooti* population density data.

1.2 Estimating Animal Abundance and Density

1.2.1 Background

A considerable amount of time and resources are devoted to preserving wildlife populations (Stokes et al., 2010). In order to do so, accurate animal abundance or density estimates are required. Animal abundance is the number of individuals (of a specific species) that occupies a site of interest (Seber, 1982). Whereas animal density (or 'relative abundance') is the number of individuals per unit area of this particular site. Animal density is critical to conservation management and is often preferred to abundance since it can be compared across space and time (Gerber et al., 2012). Especially in the case of threatened species where the population size plays an important role in deciding on the appropriate course of conservation action to be taken (Mills, 2012).

Unfortunately it is usually infeasible to count every individual over its entire range or any area of interest (Keeping, 2014). Thus, efficient monitoring techniques are necessary for the effective management of threatened species (Lettink and Armstrong, 2003). Scientists use a variety of methods to find, measure and map animal communities. Two commonly employed methods in ecology to estimate animal abundance (or density) exist: 'capture-recapture' (CR) and 'distance sampling' (DS) (Chao, 1987). The concept behind both approaches is that the probability of detecting individuals of the targeted species can be estimated and these provide information on how many individuals in the study

area were not detected and thereby allowing for the estimation of the species abundance and density (Yoccoz et al., 2001).

Distance sampling involves carrying out a survey along an array of randomly selected lines, searching for the species of interest and for each individual encountered, the perpendicular distance from the line and the individual is recorded (Thomas et al., 2014). Distance sampling then uses a 'detection function' - a mathematical function that returns the probability of detecting an individual as a function of distance from the observer (Anderson et al., 1983). A key assumption with distance sampling is that all individuals are detected if they are directly on the line. Ultimately not all individuals that the observers pass will be detected as the detection probability generally decreases with increasing distance from the surveying lines (Thomas et al., 2010). The basic idea is that distance sampling takes the factor of non-detection into account by considering that not all individuals in the transect are detected. Then it attempts to estimate the number of animals that were missed in the survey to obtain a better estimate of the species abundance (Buckland et al., 2012; Thomas et al., 2014). Animal density is the quotient of the species' abundance and the surveyed area (i.e. the ESA- 'effective sampling area'). A more detailed description of the methodology is not discussed here but can be found in *Buckland et al. (2012)*.

First applied in the estimation of the human population in London (Graunt, 1977), capture-recapture methods (also known as 'mark-recapture') were applied in the study of fish (Kipling and Cren, 1984) and wildlife populations (Gazey and Staley, 1986). Following the description given by Pollock (2000), the basic capture-recapture methods consist of a first sample of the population that is captured for marking (e.g. by using tags) and released, later another portion of the population is captured and marked once again, thereby creating a 'capture history' that shows how many times the particular individual was captured and on what occasions. Using the proportion of recaptures and individuals newly captured, it is possible to estimate animal abundance (White, 1982).

However, capture-recapture methods do not draw inferences about space explicitly and do not attempt to estimate an effective sampling area (Borchers and Marques, 2017). Basic capture-recapture models can thus only estimate abundance but not animal density (Buckland and Elston, 1993; Marques et al., 2013). In addition, basic capture-recapture methods do not account for spatial heterogeneity, whereby individuals that live or wander close to the traps are more likely to be captured. In order to account for capture heterogeneity introduced by the individuals' location in relation to traps, a further generalization of CR methods led to spatially explicit capture-recapture (SCR) methods. (Borchers and Marques, 2017; Borchers and Efford, 2008; Efford, 2004)

1.2.2 Spatially Explicit Capture-Recapture

Spatially explicit capture-recapture (SCR) methods combine capture-recapture methods with the spatial component of distance sampling *(Efford et al., 2009)*. The spatial component arises from the location of individuals where the probability of an individual being detected depends on its location - animals closer to the trap have a higher probability of being detected than those further away from the traps (Borchers, 2012). Therefore, primary SCR data - capture histories - not only include the occasions when an individual was detected but also include the spatial detection information of where the individual was detected (Borchers and Efford, 2008).

SCR accounts for spatial heterogeneity by considering that each individual has an activity centre. An individual home range is defined by "the area used by an organism during some time period" *(Efford et al., 2009)*. Within the SCR framework, activity centres are identified as the centroid of the space (i.e. home range) where these individuals spend time over the trapping periods (Efford, 2004; Royle et al., 2011). Activity centres are typically unobserved (Marques et al., 2013). The probability of capturing an individual can be modelled as a function of

the distance from those traps to the unobserved activity centres. Based on this additional information, one can specify the probability of detecting an individual by any trap on any occasion (Marques et al., 2013; Borchers and Marques, 2017; Efford et al., 2009). Assuming that activity centres are uniformly distributed in space, the integral under the detection function over space gives the effective sampling area of a given set of traps (Borchers, 2012). Thus, density estimates can be obtained directly without an ad hoc effective sampling area estimate *(Efford et al., 2009)*.

Traditionally capture history data were obtained using methods involving physical capture of individual animals only. Technological advances now provide methods to detect individuals without having to physically handle them (Borchers, 2012; Royle et al., 2013). The use of SCR methods has been spurred on with the advent of non-invasive 'proximity detectors' for sampling populations (Burton et al., 2015; Mondol et al., 2009). Rare and elusive species that historically could not be studied effectively because they were difficult or impossible to capture and physically handle can now be studied (Royle et al., 2013). Additionally, the presence of non-invasive detectors does not influence animals' behaviour. A wide range of non-invasive detection devices include camera traps (Soisalo and Cavalcanti, 2006), acoustic recording devices (Dawson and Efford, 2009; Measey et al., 2017) and molecular techniques where the DNA samples can be obtained from material dropped by animals (Mollet et al., 2015).

An appealing aspect made possible by these detectors and is that abundance and density estimates can be obtained from spatial observation accumulated over a single sampling occasion because animals are not held at traps *(Efford et al., 2009)*. The underlying idea is that redetection of individuals occurs at different points in space rather than time. SCR methods have been applied in many areas including, acoustic 'trapping' of ovenbird vocalizations (Dawson and Efford, 2009) and minke whales "boing" calls (Martin et al., 2013), camera-trapping of

wolverines (Royle et al., 2011) and genetic capture of chimpanzees faecal samples (Moore and Vigilant, 2014).

1.2.3 Acoustic Spatially Explicit Capture-Recapture

In the context of elusive species, where for example, the organism lives under water or is tiny but produce acoustically detectable and distinguishable calls, acoustic spatial capture-recapture (aSCR) methods offer an alternative survey mode (Marques et al., 2013). aSCR is a statistical method used to estimate densities of acoustically active species (Dawson and Efford, 2009; Efford et al., 2009). aSCR requires data recorded from an acoustic array; which usually consist of setting up detectors in the form of microphones (or hydrophones in marine environments; (Royle et al., 2013). Using wireless links or cables, acoustic signals are then transmitted from the widely spaced microphones to data-recording equipment (Marques et al., 2013).

The data required to estimate density using aSCR are the capture histories, created when the same animal call is detected on multiple microphones at the same time, where the recaptures are the redetections (Stevenson et al., 2015; Efford et al., 2009; Borchers, 2012). The detection and non-detection patterns at the different microphones allow the unobserved location of the source call to be estimated with an associated measurement error (Efford et al., 2009; Measey et al., 2017). With each microphone having a known location, the parameters of a detection function can be estimated whereby the function describes how detectability of a call declines with increasing distance (Borchers, 2012).

Moreover, together with capture histories, data collected from an acoustic survey contain information such as signal strength (SS) and time-of-arrival (TOA) which can be used for further inform call locations. The underlying idea behind SS and TOA is that a call arrives earlier and with a stronger signal at a microphone that is closer to the call's source. Stevenson et al. (2015) extends the conditional likelihood approach of Borchers and Efford (2008) to incorporate signal strength

and time-of-arrival information. The inclusion of those gives greater precision on the location of the sound source, better parameters of the detection function and thus call density can be estimated more accurately (Stevenson et al., 2015; Measey et al., 2017).

Assuming that sensitivity across microphones is constant, the detection function of each microphone can be used to construct a detection surface over the acoustic array. The effective sampling area can be calculated by integrating under the estimated detection surface. The detection function allows the computation of the probability of detecting a frog call at a given location by at least one microphone (Borchers, 2012; Stevenson et al., 2015). Thus, the proportion of detected calls can be calculated. A call density estimator, $\widehat{D}$, is then obtained by dividing the number of detected calls, n, by the effective sampling area and the survey length, t.

The initial acoustic SCR methodology proposed by *Efford et al. (2009)* assumed that call locations were independent of one another. However, this is unlikely to hold as individuals emitting more than one call over the course of the survey are almost certainly related. Despite the violation of this assumption, Stevenson et al. (2015) showed that although variance estimates are underestimated, the bias in point density estimates obtained from aSCR are negligible. To account for dependence between call locations and thereby correcting variance estimates, the authors made use of a parametric bootstrap (see Stevenson et al., 2015).

1.3 Research Aims and Objectives

Species distributions are often limited by environmental factors (Gaston, 2003) and according to the abundant-centre hypothesis, abundance should be highest where the environment is most favourable for the species (Péron and Altwegg, 2015; Sagarin and Gaines, 2002). With the availability of a quantitative population estimate of *A. lightfooti* across its range, the main research aim was to assess if the same environmental factors determining the occurrence of *A. lightfooti* also determine the abundance of the species inside its range. Secondary

to this, was to construct the first peninsula wide population-density surface for this visually cryptic but acoustically active species using all available information on the species to date.

The objectives of the study are as follows:

1) Fit hurdle models to occurrence data and density data in terms of different environmental variables in order to investigate if the factors determining the occurrence of *A. lightfooti* also determine the species density patterns inside its range.
2) Construct a habitat suitability map based on opportunistically collected occurrence data of *A. lightfooti* from multiple sources, using an ensemble of species distribution models.
3) Produce the first peninsula wide population density map of *A. lightfooti* using all available information on the species.

1.4 Research Outline

Chapter 2 provides an overview of the data used in this study. The details of how sites were chosen together with the surveying process are discussed. The study area is also introduced supported with satellite images that shows the locations of where the species was present and absent. Chapter 3 then provides the basic structure of the two-part count model used for modelling *A. lightfooti* density data and the different species distribution models used to model presence data.

The emerging availability of spatial variation in abundance within a species can be seen as an opportunity to re-examine ecological hypotheses once drawn from simplistic assumptions about species distribution. In Chapter 4, four hypotheses are formed and tested to evaluate the extent to which the same environmental factors limit the distribution of the species and its abundance within its distribution. Four model structures are built using hurdle models, which is a two-

part count model. The spatial distribution and abundance of the species are modelled separately to gain insight into whether the two processes are governed by the same factors.

A vast majority of species data that is available today consist of presence-only data collected under non-standardized designs. Presence records of *A. lightfooti* are available back to 1933. Those data were collected opportunistically over the years by different observers with no systematic approach. Using an ensemble of seven species distribution models, a habitat suitability map based on those opportunistically collected data and environmental variables is constructed and presented in Chapter 5. The chapter does not investigate the specific ecological interpretation of the environmental predictors but briefly evaluates the different individual species distribution models used. The resulting distribution map derived from this 'lesser quality' data is then used as baseline information to integrate with the density data of 'higher quality' in the next chapter.

In Chapter 6, I explore how the habitat suitability map constructed in chapter 5 can be used in combination with higher quality abundance data to produce the first peninsula wide population density surface of A. lightfooti. Whilst the chapter does not test any specific hypotheses, it uses all available data and a model that fits the density data well, to produce the best possible map.

Chapter 2 - Data Overview

2.1 Study Species

Arthroleptella lightfooti is a species of frog from the family of Pyxicephalidae, endemic to Table mountain and other mountains of the Cape Peninsula, South Africa (Turner and De Villiers, 2007). Adult females can reach up to 22mm in length (Figure 2.1) while the males are smaller with a short, insect like chirp advertisement call consisting of three rapid pulses of 0.1 s long at an accentuated frequency of 3.1-3.4 kHz (Channing, 2001; Measey et al., 2017). It inhabits seepages mostly in fynbos areas (Channing, 2004). Calls peak during breeding seasons between April and December (Minter, 2004). The breeding season is in the southern hemisphere winter, which is the rainy season in the south-western Cape. Females lay clutches of 5-12 eggs in wet mossy areas or under thick vegetation (Minter, 2004). This species does not occur in sympatry with any other *Arthroleptella* species (Turner and De Villiers, 2007).

Figure 2.1 - Arthroleptella lightfooti (Photo by John Measey).

2.2 Density Dataset

Acoustic SCR was first applied to call data from A. lightfooti by Borchers et al. (2015) and Stevenson et al. (2015) to assess the performance of acoustic SCR. The cryptic but acoustically detectable nature of this species makes them ideal candidates for the application of acoustic SCR methods. Moreover, the male frogs of this species remain mostly stationary while calling, thus the method need not account for any possible movement (Stevenson et al., 2015). For this study I used the data collected by Louw (2018) for an earlier study where she used aSCR to obtain density estimates of calling individuals for more than 80 sites all over the Cape peninsula and this section gives an overview of the data.

Louw (2018)'s study relied on sites where the species was present and calling, thus to make sampling more efficient, the goal was to allocate more effort to areas where the species was most likely to occur. That is, a site represented as a grid cell (at a resolution of $30m^2$), which was more likely to contain the species would be more likely to be selected for sampling. Sampling of sites was therefore stratified by habitat suitability obtained from a MaxEnt species distribution model based on 367 occurrence records.

Thus, the main sampling effort was assigned to sites where habitats were most suitable but also including sites with less suitable habitats. Sites were therefore grouped into four strata according to the habitat suitability index ranging from: 0.6 to 1, 0.4 to 0.6, 0.2 to 0.4 and 0 to 0.2. Thereafter, 80% of the sites were selected from the first stratum, 10% from the second, 5% from the third and fourth, respectively. Furthermore, a Conditioned Latin Hypercube (Minasny and McBratney, 2006) sampling, which is a stratified random sampling procedure that stratifies samples along the distribution of the following environmental gradients at those sites: slope, elevation, aspect, mean minimum winter temperature, winter solar radiation, vegetation type, Maximum Normalized Difference Vegetation

Index (NDVI) of 2015 and the most recent fire as of January 2016. It ensures that a full coverage of the range of these variables is sampled efficiently.

During the breeding season in 2016 and 2017, Louw (2018) then accessed those selected sites by walking through paths where possible and with the use of a Global Positioning System (GPS) device. Acoustic arrays were then set up where the frogs could be heard calling. If no frogs were heard at the target site, the next closest location of calling frogs was then located by walking in circle around the target site. If no frogs were heard within 100m of the target site, a density value of 0 frog m^2 was recorded for that site. Over the sampling period, 78 sites were recorded as absences. The distances between the microphones were not fixed and varied between arrays depending on the site. The immediate area around the site (200m) was vacated while recordings took place.

The calls were recorded for 40 minutes, with each of the six microphones recording on independent tracks with a sample bit rate of 48 kHz and a bit depth of 24 bits (Measey et al., 2017). When the recording was over, distances between each of the microphones were measured using a 30m measuring tape. The position of each microphone was also recorded using a GPS device. She then recorded the main features of each site (i.e. height of vegetation, presence of running water, rocks, footpaths and how the array stood in relation to these features) with sketches together with an accurate representation of every unique acoustic array that was set up in the field.

To obtain estimated calling densities of frogs at each site, the recording first needed to be processed in order to isolate the frog calls from the rest of the recorded soundscape, to determine the number of unique calls detected across the array and the microphones on which they were detected and then analysed using aSCR. This was necessary as the acoustic SCR presented by Stevenson et al. (2015) assumes that individual calls are identifiable, that is, if a call was detected several times, it is known whether or not it is the same call. She then prepared

the acoustic data (recordings) from the field for analysis using the open source software PAMGUARD (Gillespie et al., 2009).

From each recording, she then extracted the following information: the time at which a call was heard on each microphone, the signal strength of each call, and the microphones on which a call was and was not detected. The first 10 minutes of each recording was ignored as frogs typically stopped calling due to the disturbance of setting up the array. Due to the computational time and power required to obtain calling densities of animals, 10 subsamples of 1 minute were selected from the remaining 30 minutes of each recording to be run through the ascr package. Louw (2018) selected those subsamples that were free from bird calls that could be misidentified as frog calls and free from overexposure of noise that was mostly caused by wind.

Moreover, in order to convert estimated density of calls to calling animal density, an average call rate was needed. Thus, an independent recording of one-minute in length of individual calling frogs (n=13) was conducted and used to calculate an average call rate. The latter was then used as an input for aSCR analysis.

Using the `aSCR` package in R, she then obtained density estimates for more than 80 sites. A detailed description of the process can be found in Louw (2018). A parametric bootstrap was used to determine standard of errors of density estimates for each subsample within a site. Then to obtain a standard error for a site from the standard errors of its subsamples, she applied a subsampled-based modification of the bootstrap procedure described in Stevenson et al. (2015). Louw (2018)'s study then looked at the reliability of these estimates using the coefficient of variation (CV), which is a standardised measure of dispersion (Abdi, 2010). Louw (2018) considered density estimates of sites with a CV of more than 30% as less reliable population estimates.

Density estimates and observed absences are depicted on the map in Figure 2.2. The red dots represent sites at which density estimates were successfully obtained while yellow dots represent sites at which the species were absent and thus have zero density.

Figure 2.2 - Map of the Cape Peninsula, red dots representing sites for which density estimates are available and used in the study. Yellow dots represent sites are which *A. lightfooti* were absent.

2.3 Presence-only Dataset

For the presence-only dataset, a total of 494 historic occurrence points was acquired dating from 1933 to 2015. They are referred to as historic records to make it clear that they date from before the density data were sampled. These records were collected opportunistically with no record of sampling efforts. The records are spread latitudinally and longitudinally throughout the study area. The data were collated from various sources (Appendix A). Figure 2.3 below illustrates the occurrence records within the study area, where most are found within Table Mountain National Park.

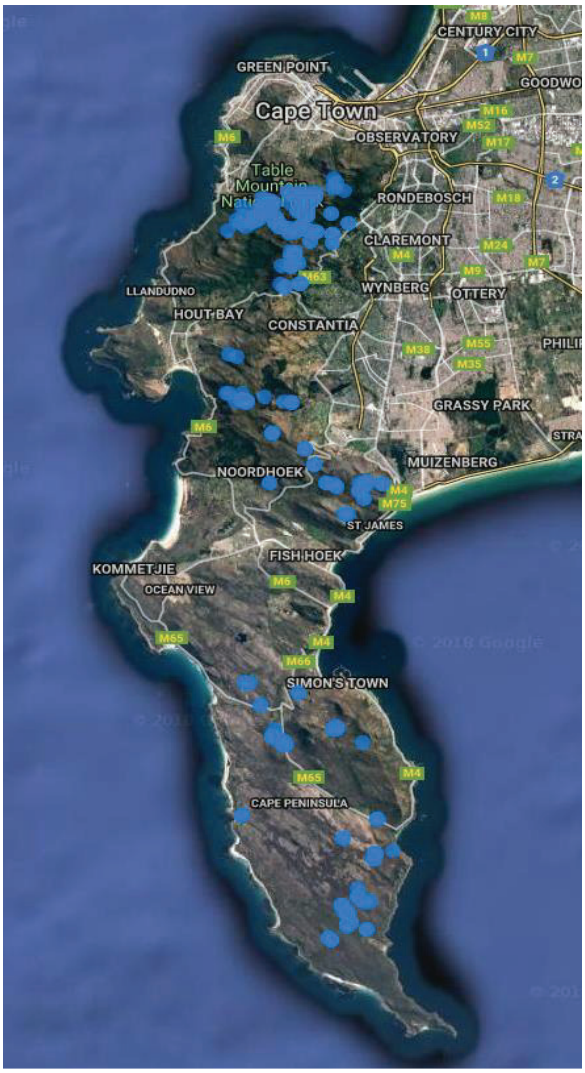


Figure 2.3 - Map of the Cape Peninsula in South Africa, where the blue dots represent occurrence presence record of *A. Lightfooti* within the study area.

2.4 Environmental Predictors

Although biotic interactions are known to affect species spatial patterns via several mechanisms (Bascompte, 2009), most species distribution models are calibrated using abiotic predictors due to the unavailability of data on biotic factors. For purpose of this study, models are built using abiotic predictors alone. The choice of the number and type of environmental variables is species-specific and relies on experts' judgement.

Aquatic and terrestrial habitats are both necessary for anurans to complete their life cycle. Amphibians are known to have poor dispersal abilities and a study by Funk et al. (2005) showed that landscapes such as mountain ridges and elevation have a particularly strong effect on the dispersal and gene flow of amphibians. Another study by de Castro Godinho and Da Silva (2018) found that the rivers in the Amazonia largely contributed in explaining the variability in anurans biogeographic regions, followed by climate and topographic variables (de Castro Godinho and Da Silva, 2018). Thus, climatic and topographical variables are the most widely used predictors in SDM and are included in this study (Miller, 2010).In addition to those, different vegetation maps and a wetland map were also included as predictors in the study. No simple explanation currently accounts for the key factors shaping anurans density.

Digital Elevation Models (DEMs) are convenient for representing varying topographic surfaces on earth (Thompson et al., 2001) and were thus used to derive the topographical variables. A 10m resolution digital elevation model was aggregated to a 30m resolution and used to derive topographic position, roughness, slope and two components of aspect: north-southness and east-westness. These layers were used to calculate the solar radiation during the warmest and coldest month using Geographic Resources Analysis Support System (GRASS).

Mean temperature surfaces were provided by the South African Environmental Observation Network (SAEON). These were obtained using spatial interpolation from 100 temperature data loggers placed across the Peninsula (Slingsby et al. Unpubl.). The wetland layer is a phenological classification based on Landsat data. The vegetation maps are the national types and subtypes (Mucina and Rutherford, 2006) and vegetation communities (Cowling et al., 2004). Table 4.2 gives a brief description of the environmental data and their sources.

Table 2.1 - Environmental variables included in the study.

Predictor Variable	Description	Source
Slope	The rate of change in elevation over a distance (continuous).	http://web1.capetown.gov.za/web1/opendataportal/AllDatasets
Elevation	Metres above mean sea level (continuous).	
East-West	East-westness = sine of aspect in degrees. Gives the direction the slope is facing in decimal fractions (continuous) from East to West.	
North-South	North-southness = cosine of aspect in degrees. Gives the direction the slope is facing in decimal fractions (continuous) from North to South.	
Roughness	Surface texture (continuous).	
TPI (Topographic Position Index)	A continuous index indicating the relative height along a hillside slope.	

Vegetation types	A categorical variable (8 levels) with different values assigned to different vegetation types.	
Vegetation subtypes	A categorical variable (8 levels) with different values assigned to different vegetation subtypes.	
Vegetation communities	A categorical variable (8 levels) with different values assigned to different vegetation communities.	
Wetland types	A categorical variable (5 levels) with different values assigned to different type of wetlands.	
Wetland Area	Also, a continuous variable giving the wetland area in each grid cell (irrespective of the wetland type).	
Summer Solar radiation	Mean maximum estimation of solar radiation during the warmest months (continuous).	https://grass.osgeo.org/grass72/manuals/r.sun.html
Winter Solar Radiation	Mean minimum estimation of solar radiation during the coldest months (continuous).	
Mean Summer temperature	Mean maximum temperature during the warmest months in °C (continuous).	South African Environmental Observation Network (SAEON)
Mean Winter temperature	Mean minimum temperature during the coldest months in °C (continuous).	

Chapter 3 – Species Abundance and Distribution Models

3.1 Introduction

Species data that describe the distribution of a species can be grouped in three broad types; presence-only data, presence-absence data and abundance data (Franklin, 2010). The class of species data available direct the types of modelling algorithm appropriate to use. This chapter reviews the different analytical tools that I will use to address the goals of this study. This primarily covers two areas:

- Species Abundance Model (SAM) – used to investigate if the factors determining the presence of *A. lightfooti* also determine the species density pattern inside its range using the density and absence data described under Section 2.2. The SAM is also used to produce the first peninsula wide population density map of *A. lightfooti.*

- Ensemble of Species Distribution Models (SDMs) – used to produce a habitat suitability map based on opportunistically collected occurrence data described under Section 2.3.

3.2 Species Abundance Models

Species Abundance Models (SAMs) are species distribution models built from count data. Response variables in ecological datasets often have a substantial proportion of zeros and a substantial positive skewness for non-zero values (Martin et al., 2005). Excess zeroes make the probability of the outcomes at zero inconsistent with baseline models such as Poisson and Negative Binomial GLMS (Ridout et al., 1998). The study of rare species often leads to the collection and analysis of data with a high frequency of zeroes (Fletcher et al., 2005; Martin et al., 2005). These types of data are often referred to as being zero-inflated. I

conducted a literature review, and several approaches have been suggested for the analysis of this kind of data:

1. Fit a generalized linear model, whereby the random variable is modelled using a Poisson or Negative Binomial distribution. However, in both these cases, the zero values and the positive values are modelled using the same distribution which often leads to a poor fit to ecological data (Welsh et al., 1996)

2. Moreover ecological data often contain more zeroes than one would expect under Poisson or Negative Binomial distribution. Lambert (1992) suggested a zero-inflated model in which these extra zeros can be modelled as an additional process. It assumes that the zeros come from two different generating processes. However, in this case, the model does not neatly separate the processes that generate the zeros and positive values.

3. Mullahy (1986) proposed a 'Hurdle Model', where one assumes a separate process that generates zeros versus the positive values and then another process determines the magnitude of the non-zero values. In particular, a hurdle model combines a dichotomous-outcomes model and a truncated count model. It is also known as a two-part model that separately models the occurrence of a zero value and the positive abundances.

The last of these approaches has been used in fields from political sciences (King, 1989) to fisheries (Stefánsson, 1996). The idea underlying the hurdle formulation is that a binomial probability model governs the binary outcome whether a count variate has a zero or a positive realization. If the realization is positive the 'hurdle' is crossed, and the conditional distribution of the positives is governed by a truncated-at-zero count data model" (Mullahy, 1986). Therefore, there are two

types of distributions where one deals with the zeros while the other one deals with the non-zero counts.

In an ecological setting, Welsh et al. (1996) called this a conditional model and suggested a Poisson or Negative Binomial distribution for the positive abundances, while Stefánsson (1996) proposed the use of a lognormal or gamma distribution for the positive values. Expression for the likelihood showing that this type of model contains two different components corresponding to the two models being fitted is given by both Stefánsson (1996) and Welsh et al. (1996).

My goal was to understand the drivers of occurrence and density of the *A. lightfooti* and one major advantage of hurdle models is that the two parts of the model can be modelled using different sets of explanatory variables, which allows researchers to gain insight into whether they are influenced by the covariates in different ways (Fletcher et al., 2005).

Thus, in this study the hurdle model is used to model the density and absence data (see Section 2.2). A GLM approach as described in section 3.3.1 is used with a logistic regression model for the first part of the model, relating the log odds with the environmental variables and a gamma regression model for the positive values.

3.2.1 The Hurdle Model

The following definition of the hurdle model is based on Winkelmann (2008). Let $f_1(.)$ be the probability mass distribution (pmf) of the binary part of the hurdle model. Let $f_2(.)$ be a pmf of the second process of it. The pmf the hurdle model is then given by:

$$f_{hurdle}(Y_i = y_i) = \begin{cases} f_1(0), & y_i = 0 \\ \left(\dfrac{1 - f_1(0)}{1 - f_2(0)}\right) f_2(y_i) \; = \; \Phi f_2(y_i), & y_i = 1,2,\dots \end{cases}$$

In the probability mass function of the hurdle-at-zero model, $(1 - f_1(0))$ is the probability that an outcome crosses the hurdle. If the outcome is positive, the hurdle is crossed with the probability $(1 - f_1(0))$. $(1 - f_2(0))$ is a normalization. It indicates the truncation of the model at zero value (Cameron and Trivedi, 2013).

The expected value of the hurdle model is determined by the probability of response outcomes at zero and by the density of the zero-truncated model (Winkelmann, 2008):

$$E[Y_i] = \frac{1 - f_1(0)}{1 - f_2(0)} \sum_{k=1}^{\infty} k\, f_2(k)$$

$$= \Phi\, (E_2[Y_i])$$

where $E_2[Y_i]$ denotes the expectation with respect to $f_2(y)$ and$\Phi = \frac{1-f_1(0)}{1-f_2(0)}$. In contrast to the zero-truncated model, the expected value of the hurdle model differs by a constant $(1 - f_1(0))/(1 - f_2(0))$ (Winkelmann, 2008; Cameron and Trivedi, 2013). In cases with excess zeroes, this constant will be less than 1. Consequently, the expected value of the hurdle model will be lower than the expected value of the baseline model (Saffari et al., 2012). In this way, the hurdle model handles the problem of over dispersion.

The variance of the hurdle model described by Cameron and Trivedi (2013) is:

$$Var[Y_i] = E\left[Y_i^2\right] - (E[Y_i])^2$$

$$= \left(\frac{1 - f_1(0)}{1 - f_2(0)}\right) \sum_{k=1}^{\infty} k^2\, f_2(k) - \left[\left(\frac{1 - f_1(0)}{1 - f_2(0)}\right) \sum_{k=1}^{\infty} k\, f_2(k)\right]^2$$

$$= \Phi E_2\left[Y_i^2\right] - [\Phi E_2[Y_i]]^2$$

The binary part of the hurdle model in this study is modelled using a binomial GLM. The probability mass function of the binomial distribution is given by:

$$f_1(y) = \frac{n!}{y!\,(n - y)!} \pi^y (1 - \pi)^{n-y}$$

For $y = 0,1,2,\dots,n$ (where n in this study is the total number of density estimates),$0 < \pi < 1$, with $E(Y) = n\pi$ and$Var(Y) = n\pi(1-\pi)$.

This thesis focuses on hurdle models with continuous data with density estimates as the response variable. The hurdle model is transformed into a Gamma Hurdle Model (GHM) by stating, $Y_i|Y_i > 0$ follows a Gamma distribution. Using the notation of Wackerly et al. (2008), $f_2(Y_i|Y_i > 0)$ is defined such that:

$$f_2(Y_i|Y_i > 0) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} {y_i}^{\alpha-1} e^{\frac{-y_i}{\beta}},$$

where $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}\,dy \;\; and \;\; y, \alpha, \beta > 0$. α is the shape parameter and β is the shape parameter. $E[Y_i|Y_i > 0] = \alpha\beta$ and $Var[Y_i|Y_i > 0] = \alpha\beta^2$.

It leads to the response variable Y_i in GHM having the distribution:

$$f_{hurdle}(y_i) = \amalg(y_i = 0)\pi + \amalg(y_i > 0)\,(1-\pi)\frac{1}{\Gamma(\alpha)\beta^{\alpha}} {y_i}^{\alpha-1} e^{\frac{-y_i}{\beta}},$$

where $\pi = P(y_i = 0)$. Adapted from Rigby and Stasinopoulos (2009)but with changes in notations, the conditional expected value and variance of the hurdle model, derived from the general expression for the mean and variance given above are as follows:

$$E[Y_i|Y_i > 0] = (1-\pi)\alpha\beta$$

$$Var[Y_i|Y_i > 0] \; = (1-\pi)\alpha\beta^2(1+\alpha\pi)$$

3.2.2 Assessing the performance of Hurdle Models

As a measure of the performance of the hurdle model, in this study the AIC value was used. Akaike Information Criterion (AIC, Akaike, 1974) is a common criterion used for model evaluation. It works by penalizing model complexity, in the aim to rank models according to how well they balance the risks of overfitting

and underfitting (Hu, 2007). The AIC is defined as follows (Symonds and Moussalli, 2011):

$$AIC = -2logL + 2p$$

Here, $logL$ is the maximised log-likelihood for a particular model and p is the number of parameters in the model. The first term is a measure of fit; a measure of how closely the model fits the data. The second term is a penalty term because as more parameters are used, the better the model will be able to explain the data, even if the parameters are not related (Symonds and Moussalli, 2011). The AIC measure the relative goodness of fit of the models, but does not provide a measure of absolute goodness of fit (Langrock et al., 2012). In other words, if all models fit poorly, the model which fits the least poorly will be selected. The model with the smallest AIC value in the set is therefore the best model in the set (Hu, 2007). In this case the hurdle model with the smallest AIC value was considered as the best model.

Other criteria that are used for model comparison are the Bayesian Information Criteria (BIC), which usually select the models with fewer parameters than the AIC (Sawa, 1978) and the Integrated Complete Likelihood (ICL) criterion proposed by Biernacki et al. (2000).

3.3 Ensemble SDM

Recent decades have witnessed a burst of interest in species distribution modelling (Franklin, 2010; Guisan et al., 2017). This has resulted from significant threats facing ecological systems, such as climate change and habitat destruction (Hefley and Hooten, 2016), which led to a growing need for describing, understanding and predicting the geographical distribution of biodiversity. Advances in Geographic Information Systems (GIS), Global Positioning System (GPS) and in statistical methods have given ecologists the opportunity to comprehend and estimate the species distributional areas based on the relation of

known species occurrences and environmental variables (Maggini, 2011; Peterson and Soberón, 2012).

SDMs relate the pattern of presences and absences to a set of variables thought to limit a species' distribution. Then, these relationships are used to project the species distribution in geographic space (Elith and Leathwick, 2009). The relationship that the SDMs describe is considered as a representation of the species niche. However, the niche concept is quite nebulous in the area of species distribution modelling (Soberón, 2007).

SDMs are built using a variety of statistical methods which vary in complexity as different models assume different occurrence-environmental relationship (Merow et al., 2013b). The choice of model generally depends on the type and quality of biological and environmental data available (Elith and Graham, 2009). Figure 3.1 below presents a typology of well-known species distribution models (Guisan and Zimmermann, 2000). With the different statistical algorithms available for modelling, recent studies have shown that disparities among different models structures can be very large, making model selection difficult (Guisan et al., 2017; Marmion et al., 2009). An alternative is to use an ensemble of models.

The concept of ensemble modelling is to avoid selecting one single best model but instead to use a group of algorithms for inference (Marmion et al., 2009). The different resulting 'habitat suitability indexes' are then combined to get a single value per site (or 'grid cell'). In other words, the presence of a species might be well classified by some models and misclassified by others. Thus by making use of an ensemble model, one can reduce the predictive uncertainty of a single model by combining predictions (Grenouillet et al., 2011; Marmion et al., 2009).

Marmion et al. (2009) investigated five different techniques of combining several modelling algorithms – mean all, median all, weighted average, median principal component analysis and 'best'. They identified the 'mean all' consensus method as one of the best approaches to building an ensemble SDM, whereby the arithmetic

mean of the predictions made by each model for a particular grid cell is computed to form the ensemble prediction map. A more detailed description of the other approaches can be found in Marmion et al. (2009). Moreover in order to reduce the variability between predictions among different SDMs, Capinha and Anastácio (2011) used an ensemble of species distribution models in the analysis of four invasive decapods species by averaging the predictions from eight individual models.

Thus, in this study, an ensemble model is built from the opportunistically presence-only data (see Section 2.3) by computing the mean value of the whole predictions from 7 single SDMs. The models listed under grey backgrounds are algorithms used in this study and are described accordingly.

Species Data for Species Distribution Models

Presence and absence data consists of coordinates of the locations where the species of interest has been recorded and locations where the species has not been found (Cáceres and Legendre, 2009). A clear majority of species data that is available today consist of presence-only data (Azzurro et al., 2013) such as the opportunistically collected *A. lightfooti* presence records described in section 2.3. It is often the case due to time, financial constraints as well as to data collection procedures aiming at inventories instead of statistical analysis (Zaniewski et al., 2002).

However, in the unavailability of true absences (locations where a species is recorded as being absent), pseudo-absences can be alternatively used (Franklin, 2010). Pseudo-absences are observations sampled from all locations where a presence was not observed (Phillips et al., 2009). Other than random selection, there are different ways of selecting pseudo-absences. Some studies have used weighted survey designs which assign bigger weights to locations

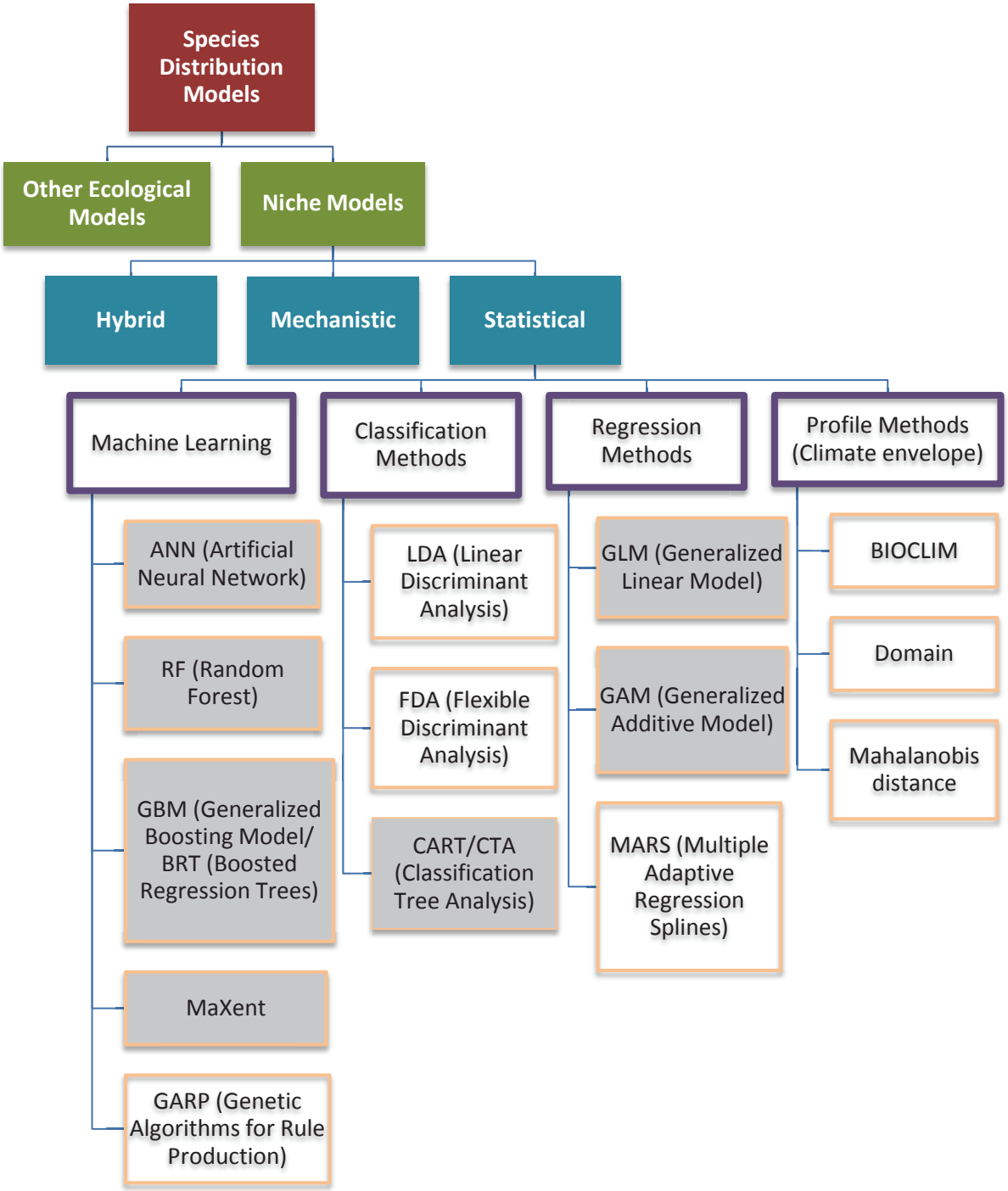


Figure 3.1 - SDM tree presenting a non-exhaustive list of species distribution models. Models listed under a grey background are algorithms used in the study.

where it is most unlikely to observe the species. Lütolf et al. (2006) excluded locations where a species was documented to have historically existed from possible background locations. It is important to generate pseudo-absences as accurately as possible to correctly classify the conditions of absence locations which can lead to more accurate species distribution models (Chefaoui and Lobo, 2008; Franklin, 2010).

For purpose of this study, I followed guidelines proposed by Barbet-Massin et al. (2012) on how many pseudo-absences to be generated based on simulated species distributions. In the study they made use of seven different SDMs from regression techniques to classification and machine learning models and investigated the models' performance based on different sample sizes of presence and pseudo-absence points. The authors found that different SDM behaved differently regarding the different sample sizes used. Regression techniques performed best when a large number of pseudo-absences were used while for classification and machine learning techniques, the models' predictive accuracy were higher when few pseudo-absences or not more than the number of presences were used (Barbet-Massin et al., 2012).

3.3.1 Generalized Linear Models (GLMs)

In linear regression model, the aim is to explain how a response variable (also referred to as dependent variable), Y, relates to one or more predictor variables (also called explanatory variables, independent variables or covariates), $X = (X_1, X_2, \dots, X_p)$:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

where $\hat{\beta}$ is the vector of estimated coefficients, and $\hat{\beta}_0$ is an estimated constant known as the intercept. It is assumed that the error term, ε, is normally distributed with zero mean and constant variance, and the variance of Y is constant across observations (Franklin, 2010). Often these assumptions are not met when dealing with ecological data (Guisan et al., 2002).

Generalized linear models (GLMs) are an extension of linear regression models that allow for non-constant variance structures in response variable (McCullagh and Nelder, 1989), for example, binary responses where there are only two possible outcomes. In the case where the response variable does not have a normal distribution, instead of transforming the response variable, GLMs uses a link function and then applies a linear model (McCullagh and Nelder, 1989)

As many data in ecology are not Gaussian, there are other distributions that can be used to characterize the different type of response variable, such as Poisson, binomial, negative binomial and gamma (Guisan and Zimmermann, 2000). These are collectively referred to as the exponential family of distributions (Nelder and Wedderburn, 1972). For example, in the case where the response variable is binary (it can assume two states, dead or alive), it may be described by a binomial distribution.

The linear model is then generalized using a link function which provides a transformation of the response variable, so that the transformed response is linearly related to the predictors and a variance function that defines how the variance of the response variable depends on its mean. Following notation from Franklin 2010, the generalized linear model can be expressed as:

$$\mathbf{G}\left(E(Y)\right) = LP = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j + \varepsilon$$

where the predictor variables $X_j (j = 1, \ldots, p)$ are combined to produce a linear predictor LP which is related to the expected value E(Y) of the response variable Y through a link function G(). Thus, formulating a generalized linear model for

SDM involves selecting a distribution for the response variable, a link function (together called the family of the GLM), the variance function and the predictor variables (McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972).

The link function defines the form of the relationship between the expected value of the response and the explanatory variables. The choice of the link function depends on the type of data. For example, if the response is in binary form, then a binomial distribution is used to characterize the response Y and the logit link is used (McCullagh and Nelder, 1989). In cases where the response is count data, a log link function is used. Species occurrence data often consist of presence and absence observations which is in binary form, consequently the logit link function is widely used in species distribution modelling (Pollock et al., 2014). Following notations from Franklin (2010), the logit link is:

$$LP = logit(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

where $\left(\frac{\mu}{1-\mu}\right)$ is the odds of success, $\mu = E(Y)$ is the probability of class 1 and $(1-\mu)$ is the probability of class 0. A log-odd of 0 means that both classes are equally likely. A GLM with a logit link is often called logistic regression (Hosmer Jr et al., 2013). The coefficients $(\hat{\beta})$ of the predictors in a logistic regression model estimate the change in the log-odds of success per unit increase in the predictor variable (Bewick et al., 2005). $\mathrm{Exp}(\hat{\beta}_j)$ estimates the factor by which the odds (of success) change per unit increase in the predictor variable. Predicted values from the LP are on the logit scale, so that an inverse transformation is needed to convert those logs of odds back to a probability. This can be done by inverting the above formula and solving for μ (Franklin 2010).

$$\mu = \left(\frac{e^{LP}}{1+e^{LP}}\right)$$

The variance function, $V(\mu)$, describes how the variance of Y depends on the mean, giving:

$$V(Y) = \phi V(\mu),$$

where ϕ is a scale (also known as dispersion) parameter.

Moreover, non-linear responses can be modelled by including polynomial or other parametric transformations (Balakrishnan, 1991). The choice for the appropriate transformation can be identified through scatter plots of partial residuals.

Parameter Estimation

Various techniques are used to estimate the parameters of GLMs (Bolker et al., 2009), but for this study Maximum Likelihood estimation is used. Scholz (2014) defines the likelihood function as the probability of the observed data given the assumed model, as a function of unknown parameters. Maximum likelihood estimation involves choosing those values of the parameters that maximize the likelihood (or equivalently, the log-likelihood), that is, the values which maximize the probability of occurrence of the observations are chosen (Pan and Fang, 2002).

3.3.2 Generalized Additive Models (GAMs)

Generalized additive models developed by Hastie and Tibshirani (1990) are a non-parametric extension of GLMs with a smoothing function. GAMs allow for an automated approach to identifying and describing non-linear relationships between predictors and response. GAMs are similar to GLMs as in both cases the probability distribution characterizing the type of response variable must be specified (Guisan et al., 2002; Miller, 2010).

The formulation for the GAM can be written:

$$g\big(E(Y)\big) = LP = \hat{\beta}_0 + \sum_{j=1}^{p} f_j\,(X_j)$$

where the coefficients of the GLM are replaced with non-parametric smoothing functions of the predictor variables, $f_j(X_j)$. The functions' shape is estimated

directly from the data. Generally it is a scatterplot smoothing function that attempts to capture important patterns in the data (Franklin, 2010; Hastie and Tibshirani, 1990). A number of scatterplot smoothers are available in GAMs, though here we focus on cubic smoothing splines.

Cubic smoothing splines are the most commonly used, whereby the ordered dataset is divided at regular intervals of values, often referred as 'knots' (Guisan et al., 2017). The intervals between those knots are then represented by a collection of polynomials of degree no more than 3. Nevertheless, the number of knots and the degree of smoothing applied when fitting the curves need to be specified by the user or can be selected through cross-validation. GAMs (i.e. the smoothing function) can be fitted using the back fitting algorithm proposed by Hastie and Tibshirani (1990) which involves the iterative smoothing of partial residuals, with respect to the covariates that the smooth relates to (Hastie and Tibshirani, 1990). A more detailed description of the process can be found in Wood (2006).

3.3.3 Classification Tree Analysis (CTA)

Classification Tree Analysis also referred to as Classification and Regression Trees (CaRT) first introduced by Breiman et al. (1984), are tree-based model ideally suited for modelling ecological data that show complex and unbalanced characteristics. The basic idea behind this method is to repeatedly split the data into groups, so that observations that end up in the same group have similar responses (Breiman, 2017). Regression trees are used for numerical responses and classification tree for categorical responses respectively.

The data are repeatedly split into mutually exclusive homogenous groups whereby each split (node) is defined by a simple rule based on a single explanatory variable. The rule for dividing the data is usually based on minimizing the classification error rate. The procedure is then applied to each individual group separately. The aim is to divide the response into homogenous groups but at the

same time keep the tree fairly small to avoid overfitting. The way that explanatory variables are used to form splits depend on their type. Splits that maximise the homogeneity of the two resulting groups are selected. The size of a tree is then determined by the number of final groups (Breiman, 2017).

The approach that has proven to be most effective in building trees according to Breiman et al. (1984) is to 'grow' a large tree using a fairly liberal stopping criterion, for example, when the number of observations in each terminal node would fall below an established minimum. Then 'prune' the tree, that is, dropping internal nodes that contribute less towards the groups' homogeneity. Simpler trees are easier to understand and are less likely to over fit the data.

3.3.4 Random Forest (RF)

Random Forests (RFs) are an extension of Classification Trees proposed by Breiman (2001) in which a large number of de-correlated trees are built with random subsets of data but also each split in each tree is developed with a random subset of predictor variables. The trees are built to a maximum size without pruning and the resulting predictions are then averaged (Cutler et al., 2007; Franklin, 2010). By averaging over several trees, there is thus a significant lower risk of overfitting.

Cutler et al. (2007) review the procedure beginning with the selection of many bootstrapped sample of the data, where all random subsets have the same number of data points and each data points has the same probability of being selected. Then a classification tree is fitted to each bootstrapped sample. However, trees produced using bootstrapped samples are highly correlated, thus each node is split using a random subset of predictor variables. It thus reduces the bias of having a very significant covariate that would make all trees look similar.

Part of the data used to calibrate the model is called the 'training' data and part of the data that is not used to build trees is referred to as 'out-of-bag' data

(Cutler et al., 2007; Liaw and Wiener, 2002). Then the fully-grown trees are used to predict those 'out-of-bag' data points and to estimate model errors and accuracies (Cutler et al., 2007; Franklin, 2010).

3.3.5 Generalized Boosting Models (GBMs)

Also known as Boosted Regression Trees (BRTs), GBMs combines the strength of two algorithms: Regression trees and boosting (Elith et al., 2008). Like Random Forests, the general idea is to fit repeatedly many simple trees to improve the accuracy of the model (Guisan et al., 2017). One difference among those two methods is the way data is selected to build the trees.

Using Random forest, each data point has an equal probability of being selected in subsequent samples. GBMs uses boosting, whereby very simple trees are built and for each new tree the input data are weighted in such a way that data that were poorly modelled by previous trees has a higher probability of being selected in the new tree (Elith et al., 2008; Franklin, 2010). The model continuously tries to improve its accuracy by sequentially fitting trees to the training data and taking into considering errors from previous trees that are built (De'Ath, 2007).

Elith et al. (2008) describes two parameters that need to be specified by users:

- Tree complexity: this controls the number of splits in each tree.
- Learning rate: also known as the shrinkage parameter, determines the contribution of each tree to the growing model.

GBMs are robust algorithms that work well with large datasets. They can be used with a variety of response types (Franklin, 2010).

3.3.6 Artificial Neural Networks (ANNs)

The concept of Artificial neural networks were first developed in 1943 (McCulloch and Pitts, 1943), inspired by biological neural networks, in particular the human brain (Franklin, 2010). ANN is another machine learning algorithm, which just like the brain consists of a large number of connections and nodes instead of neurons (Abraham, 2005). Neural networks refer to a wide collection of models, but the most extensively used in ecology is the single hidden layer back-propagation network which is described here in this section (Lek and Guégan, 1999).

These artificial neurons or nodes are organised in layers: an input, a hidden and an output layer, all connected by weights (Abraham, 2005). These weights are a result of the relationship strength between layers and their components. The input layer consists of the predictor variables where each node represents one environmental variable. The nodes in the input layer are connected to one or more nodes in the hidden layer through weighted connections based on their importance. These weights, initially randomised, are adjusted and optimized by the model in subsequent runs through the back-propagation algorithm (Abraham, 2005; Gevrey et al., 2003; Lek and Guégan, 1999).

Each node in the hidden layer consists of a combination of predictor variables which have been multiplied by the weight of the connection and summed. The weighted sums in each of the hidden layer nodes are scaled by a non-linear 'activation function'. Similarly, the connections between the hidden layer and output are also weighted. Thus, the response variable or output in neural network terminology is the result of the weighted sum of the hidden nodes. This is done a number of times and in several blocks, which give different outputs, because the weights are initially randomized.

The neural network learns the right weights by computing the error between the model's prediction and the target outcomes through the back-propagation process.

A more detailed description of the latter can be found in Riedmiller and Braun (1993). A pass through the back-propagation process with the weights updated after each pass is referred to as a cycle or epoch. Training is carried out repeatedly with weights being updated at the end of every cycle, until the model converges (Goh, 1995).

In this study, to implement the algorithm, two parameters are specified:

- The number of units in the hidden layer (size).
- The weight decay; which serve as a penalty for complexity and avoid overfitting.

The optimal size of the hidden layer and weight decay are selected using a cross-validation procedure.

3.3.7 Maximum Entropy (MaxEnt)

First introduced by Phillips et al. (2004), MaxEnt is a machine learning technique that estimates from presence data, a species' geographic distribution by finding the probability distribution with maximum entropy subject to a set of constraints (Phillips et al., 2006).

Theoretically, MaxEnt differentiates between observed presence data $(y = 1)$ to the available environment within the study area. Following Elith et al.2011, we denote z as a vector of environmental variables and background as the all locations within the landscape of interest. $f(z)$ is thus defined as the probability density that characterizes the available environment within the study area and $f_1(z)$ as the probability density that characterizes the environment across locations within the study area, where the species occurs. MaxEnt then estimates the ratio $f_1(z)/f(z)$ which gives the relative environmental suitability for presence of a species for each location within the landscape.

It does this by generating all possible distributions of $f_1(z)$ but chooses the one that maximises the similarity to $f(z)$. This distance from $f(z)$ is considered as the relative entropy of $f_1(z)$ with respect to $f(z)$. In fact, $f(z)$ is seen as a null model for $f_1(z)$, as without presence data there would be no basis to expect species to prefer any specific environmental condition. Thus, one would predict that the species occupies environmental conditions proportionately to their availability within the landscape of interest.

Moreover, the relationship between $f_1(z)/f(z)$ and the environmental variables are subject to a set of constraints. Constraints are enforced to ensure that the result is one that reflects the characteristics of the locations where the species has been observed. Taking winter temperature as an example, then constraints ensures that the mean winter temperature of the estimated probability density $f_1(z)$ is close to its mean across locations where the species was observed.

Similar to regression-based approaches where splines, linear or quadratic terms are used to better fit the relationship between occurrence data and environmental variables, MaxEnt offers different ways of modelling the relationship between $f_1(z)/f(z)$ and the environmental variables. These transformations of the environmental variables are known as features, which here also allows complex relationships to be modelled. There are six feature classes: linear, quadratic, hinge, product, threshold and categorical (see Elith et al., 2011).

To avoid overfitting MaxEnt makes use of a regularization parameter for each feature which estimates how close the expected value should be to the observed value. In other words, instead of fitting the model using exact constraints of the environmental variable, such as the mean winter temperature, it takes into consideration the confidence interval around the constraints. Rather than being treated like goals, constraints only have to be respected.

3.3.8 Assessing model performances

If independent data are not available, a commonly used approach in species distribution modelling is to partition the data into one portion used to fit the model and the other portion to validate the predictions (Guisan and Zimmermann, 2000). Therefore, in this study, a *k*-fold cross validation technique is used.

As described by Stockwell (1992) the *k*-fold cross validation procedure consists of: (1) The data set is partitioned into *k* equally sized randomly selected subsamples, (2) one subsample is excluded and the model is trained with the other *k*-1 subsamples. (3) The model is then tested on the excluded subsample, (4) the steps two to three are then repeated, excluding a different subsample each time. (5) The estimate of the model accuracy is then computed from the mean of the tested samples. What to do with these testing data?

To evaluate and compare the performance of the different species distribution models used in this study, each model was assessed using the area under the receiver operating characteristic (ROC) curve, known as the AUC (Peterson et al., 2008).

Most species distribution modelling methods return a probability of species occurrence (Franklin, 2010). The models input variable is a binary presence/absence indicator, thus to compare those model predictions to test data, one needs to transform those continuous predictions to a categorical one (presence/absence) based on a threshold probability value (Manel et al., 2001). The threshold probability is the probability value above which a prediction is considered to be positive (Franklin, 2010). The comparison between the categorical predictions with the test dataset can be summarised in a confusion matrix. However, the conversion from the probability of species occurrence to a binary output depends on a threshold which is based on subjectivity.

The area under the receiver operating characteristic curve is a threshold independent measure and is thus used in this study. The ROC curve plots the proportion of true positives against the proportion of false positives (false identification) across all possible thresholds between 0 and 1 (Jiménez-Valverde, 2012; Lobo et al., 2008). The AUC is then computed by summing the area under the ROC curve, which ranges from 0.5 to 1 (Franklin, 2010). The closer the ROC curve follows the y-axis, the larger the area under the curve and thus the more accurate the model. A model is considered to perform better than random when the AUC is higher than 0.5 (Hanley and McNeil, 1982; Roura-Pascual et al., 2009; Soberon and Peterson, 2005).

3.4 Key SDM and SAM Assumptions

One essential assumption described by Elith and Leathwick (2009) that needed to be taken into consideration before making any inferences, are that species are at equilibrium with their environment. In other words, the suitable habitat of the species is fully occupied. Although this is a required assumption for projecting the model in space or time, a few critical considerations have been raised in the recent literature on how close a given modelled system is to an equilibrium (Araújo et al., 2005). In many cases, a species is not found in an area with potentially suitable habitat because of dispersal limitation and species distribution or abundance models tend to ignore such parameter.

Chapter 4 – Are the same factors affecting the distribution of *A. lightfooti* also affecting the density where it occurs?

4.1 Introduction

It has been long recognized that the abundance and distribution of a species are closely interrelated (Brown, 1984). Species distributions are often limited by environmental factors and several studies suggest that abundance should be highest where the environment is most favourable for the species (Ehrlén and Morris, 2015; Hanski, 1993; Heino, 2005). But, do the same environmental factors determine presence and abundance patterns inside the range? I examined this question using statistically rigorous estimates of the *Arthroleptella lightfooti* population density across its entire range.

Species distribution models have largely been used to describe species occurrence patterns whereas abundance models are less common, often due to the lack of data on spatial variation in abundance within the species distribution (Sagarin et al., 2006). Consequently inferences on species habitat preferences were originally drawn on simplistic assumptions about species distribution (Sagarin et al., 2006).

However, two distinct characteristics of abundance data in ecology are their tendency to contain many zero values (Howard et al., 2014) and having a skewed distribution (Fletcher et al., 2005). When the number of zeroes, that is absences, is very large and the data do not follow standard statistical distributions, such dataset are said to be zero-inflated (Heilbron, 1994). For example, the study of rare organisms will often lead to the collection and analysis of data with a high number of zeroes (Welsh et al., 1996). There are several approaches developed to tackle the problem of excessive number of zeroes. Cragg (1971) proposed the hurdle model, which separately models the occurrence of a zero value (i.e. absence in this case) and the positive abundances (Buntin and Zaslavsky, 2004).

Using hurdle models, the aim of the study is to gain insight into whether the two processes, that is, the distribution and relative abundance patterns of the *A. lightfooti* are governed by the same environmental factors. I conducted a literature review and found that in Africa, changes in amphibian densities were mostly attributed to habitat and climate change (Hirschfeld et al., 2016; Measey and Tolley, 2011). Although inhabiting a largely protected area, *A. lightfooti* also faces these major threats, including the invasion of alien plants and frequent and intense fires (Measey and Tolley, 2011). Taking these findings into consideration, the next section will present the hypotheses developed and provide the research model that will be used in the study.

4.2 Hypotheses and Research model

With the increasing number of research being conducted, it is hard to argue that environmental gradients do not play an important role in species distribution and abundance. However, information pertaining to rare or cryptic species is not easily available and thus making it hard to examine the relation between these species and the environment. For example, information pertaining to its population density of the Cape Peninsula moss frog (*A. lightfooti*) - were previously unavailable before the advent of the aSCR (see Section 1.2.3). The hypotheses developed in this section are therefore derived from what is known to affect amphibians' distribution and abundance in order to fit that research gap apropos the *A. lightfooti* population density. The five hypotheses and research model are listed below (a description of the variables used can be found in section 2.4):

A: Climate. One of the causes of amphibians declines is related to climate change (Beebee and Griffiths, 2005; Corn, 2005). Amphibians are ectotherms, meaning that their body temperature is determined by the temperature of the environment (Raske et al., 2012). It is known that southern African frogs are most active when the surrounding temperature is 20-30°C (Navas et al., 2013; Loveridge, 1976;

Channing, 2004). Studies carried out by Measey et al. (2017) on the calling behaviour of the male *A.lightfooti* suggest that these small moss frogs showed strong seasonality in calling ecology. The breeding season is in the southern hemisphere winter, which is the rainy season in the south-western Cape. Using the *A. lightfooti* population density we investigate if high solar radiation during winter influences the population density of this species where it occurs. Moreover, we examine if there were an optimum temperature preferred by the species by including a quadratic term for temperature. We also investigate if these factors have the same influence on its distribution.

$$\text{Structure} \sim \text{Minimum Winter Temperature} + (\text{Minimum Winter Temperature})^2 + \text{Winter Solar Radiation}$$

B. Topography. The *A. lightfooti* moss frog is known to be distributed across the Cape Peninsula and occurs in variable forms of terrain (Minter, 2004). Suitable habitat for this species is often widely dispersed in a large matrix of unsuitable habitat (Turner and Channing, 2017). We thus investigate if the factors (such as slope and slope-Aspect) affecting its distribution have the same impact on its density.

$$\text{Structure} \sim \text{EastWest} + \text{NorthSouth} + \text{Slope}$$

C. Vegetation. The *A. lightfooti* are mostly associated with moss which can be distinguished under certain vegetation types (Measey et al., 2017; Du Preez, 2015), with the available population densities we can now investigate how the different types of vegetation mostly found across the Cape Peninsula influences the population densities of the species where is occurs.

$$\text{Structure} \sim \text{I(Hangklip Sand Fynbos)} + \text{I(Peninsula Sandstone Fynbos)} + \text{I(Other)}$$

Where I(.) indicates an indicator function.

D. Wetland. This species of frog is not found in open water but is mostly associated with moist terrestrial micro-habitats such as seepages, both in open fynbos and steep slopes (Channing, 2004). This small ectotherm is especially vulnerable to dehydration and therefore requires moist habitats (Channing, 2004). With breeding commencing with winter rains, the lightfoot's moss frog tadpoles develop in damp terrestrial nest (Channing, 2001; Du Preez, 2015; Measey et al., 2017).Thus we would expect higher population densities associated with large wetland areas (e.g. hill's slope seepages and floodplains). Thus, we want to investigate if the proportion of wetland present where the species occur affects their population density.

Structure ~ Wetland area

E. Null Model. The study aims at looking at the different factors mentioned above together with the possibility of perhaps having other factors influencing the distribution and abundance of those species. As a comparison to other models that do include independent variables, a null model is included (Harvey et al., 1983). I included a null model which looks at patterns that might arise through chance and in absence of ecological processes that have not been thought of or have any available environmental data.

Structure ~ 1

4.3 The Data

The data used in the chapter consist of the density data obtained from Louw (2018) and a detailed description of the data is given in Section 2.2. I was interested in understanding how and why density varies in space. Thus, sites that were visited more than once were omitted and only the most recent estimates were kept for analysis giving a total of 167 sites, each visited once. 89 of those

sites have an associated density estimate and standard error, while at the remaining 78 sites, the species was observed as being absent.

The subset of data considered in the study in order to address the ecological questions consists of eight environmental variables (see section 2.4), grouped under the four hypotheses mentioned in the previous section. Environmental data at each site were extracted from a 30x30m grid-map. The response variable is the estimated population density of the lightfoot's moss frog (individuals/m^2). The explanatory variables were: mean minimum winter temperature, winter solar radiation, slope aspect (eastwest and northsouth), slope, vegetation types (hangklip sand fynbos, Peninsula sandstone fynbos and others) and wetland area (m^2).

4.4 The Hurdle Model Analysis

The analysis consisted of three stages. The first involved dividing the 167 data points into two datasets: one indicating whether *A. lightfooti* was present or not at each site, which therefore consists of 89 presences and 78 absences represented by 1s and 0s. The second dataset contains the population density estimates together with the associated standard errors for those sites where the frog was present. These two datasets are referred to as 'presence/absence data' and the 'density data', respectively. The density dataset contained fewer observations than the presence data, as it excluded those sites where *A. lightfooti* was absent.

Before fitting the hurdle model to the data, appropriate distribution to model the continuous positive density data need to be found. The lognormal and gamma distributions have been predominantly used to model continuous positive data since they rise sharply around zero and are strictly positive (Young and Young, 2013) and are therefore considered in the study. All statistical procedures were conducted using R, version 3.5.1 (R Development Core Team, 2017). Each of the parametric distributions were fitted to the density dataset, one at a time using the

fitdistr function in the R MASS package (Ripley et al., 2013), in order to elect the distribution giving the closest fit to the data. The distribution parameters are estimated using the maximum likelihood approach. The distributions were then evaluated graphically and compared using Akaike Information Criterion (AIC).

Correlation values were determined between all continuous predictor variables using Spearman's Rho (Gautheir, 2001). Severe collinearity among predictors causes coefficient estimates to be highly unstable, making the estimates very sensitive to minor changes in the model and thus difficult to interpret. The output from a Spearman's test ranges from 1 to -1, where a value of one indicates a perfect positive correlation between two variables, a value of negative one indicates a perfect negative correlation and a value of zero indicates that no discernible relationship exists.

The second stage consisted of fitting a hurdle model to the data where the presence/absence data is modelled using logistic regression and the density data is separately modelled using gamma regression. To investigate the five hypotheses mentioned in section 3.2, five model structures were formulated where both the presence/absence data and the density data were modelled in terms of the explanatory variables. The analysis in this chapter is meant to test these a-priori hypotheses and not necessarily find a model that fits this particular dataset best.

For the presence data, logistic regression models from the R package stats was used to fit the models adopting the logit link function. The presence data being the response variable coded as 1 for a presence and 0 for an absence and the total sample size was 189. Secondly, the mean density of calling animals (i.e. the dependent variable) was modelled as having a gamma distribution, using a log link function to relate the mean to a linear combination of the predictors, which in turn varied for each of the five models presented in section 4.2 (e.g. mean minimum winter temperature and winter solar radiation for the climatic model).

The sample size used for the density data analysis consisted of 89 density estimates.

Moreover, data points that are estimated from an auxiliary analysis are often accompanied with a degree of uncertainty. Modelling these data points without propagating the uncertainty, one is assuming that each data point provides equally precise information about the deterministic part of the process variation (Goodchild, 1993; Herron, 1999). In other words, one is assuming that the standard deviation of the error term is constant over all values of the predictor variable leading to deceptive estimates that overstates precision. Thus, in order to account for the uncertainties associated with the density estimates in this study, a weight was assigned to each observation with values in 'weights' being inversely proportional to the coefficient of variation[1]. Thus, an estimate with a large standard of error will be assigned a smaller weight when fitting the model. A total of five logistic regression models and five log-linear models were fitted, representing the five hypotheses outlined in section 4.2.

The models' performances were then compared using AIC. In the case of hurdle models, where the two mechanisms are assumed to be conditionally independent, the AIC of the models are obtained by computing the sum of the AIC value of the two separate independent model parts, that is, models constructed using presence/absence data and density data (refer to section 3.2). The best hurdle model is the one with the lowest gross AIC.

The third stage involved combining the final best models for the presence and abundance data to predict the expected density of the Lightfoot's moss frog as follows. Let Y be the density of *A. lightfooti* and Z be a binary variable, equal to one when the frog is present and zero otherwise. The expected value of Y at a specific site is given by (derived from section 3.2.1):

[1] Each calling animal density estimate had an associated standard error which was used to calculate a coefficient of variance.

$$E[Y] = P(Z = 1)E(Y|Z = 1) + P(Z = 0)E(Y|Z = 0)$$
$$= P(Z = 1)E(Y|Z = 1)$$
$$= \pi\mu,$$

where $\pi = P(Z = 1)$ and $\mu = E(Y|Z = 1)$. A natural estimate of the expected density of ***A. lightfooti*** is given by:

$$\widehat{E(Y)} = \hat{\pi}\hat{\mu}$$

$\hat{\pi}$ and $\hat{\mu}$ are the estimates of π and μ obtained from the two regression models.

Advances in statistical methodology allow the construction of highly accurate approximate confidence intervals, even for very complicated probability models and intricate data structures (DiCiccio and Efron, 1996). Thus, in this thesis, a confidence interval for the expected population density of the lightfoot's moss frog was obtained using non-parametric bootstrapping (Davison and Hinkley, 1997).

The procedure involves resampling with replacement from the original sample data, fitting the best hurdle model and hence generating 'n' alternative values of $\widehat{E(Y)}$. Repeating this B times – we get B vectors of length 'n' stored in a matrix. We then sorted each vector of the predicted values and then got the quantiles for the confidence interval. There are a number of options for using the bootstrap sample of values $\widehat{E(Y)}$ to produce confidence interval (Davison and Hinkley, 1997): we choose to construct a Basic Percentile 95% confidence interval from 1000 bootstrap samples.

4.5 Results

The gamma distribution provided a better fit to the data on lightfoot's moss frog densities than the lognormal distribution (Figure 4.1).

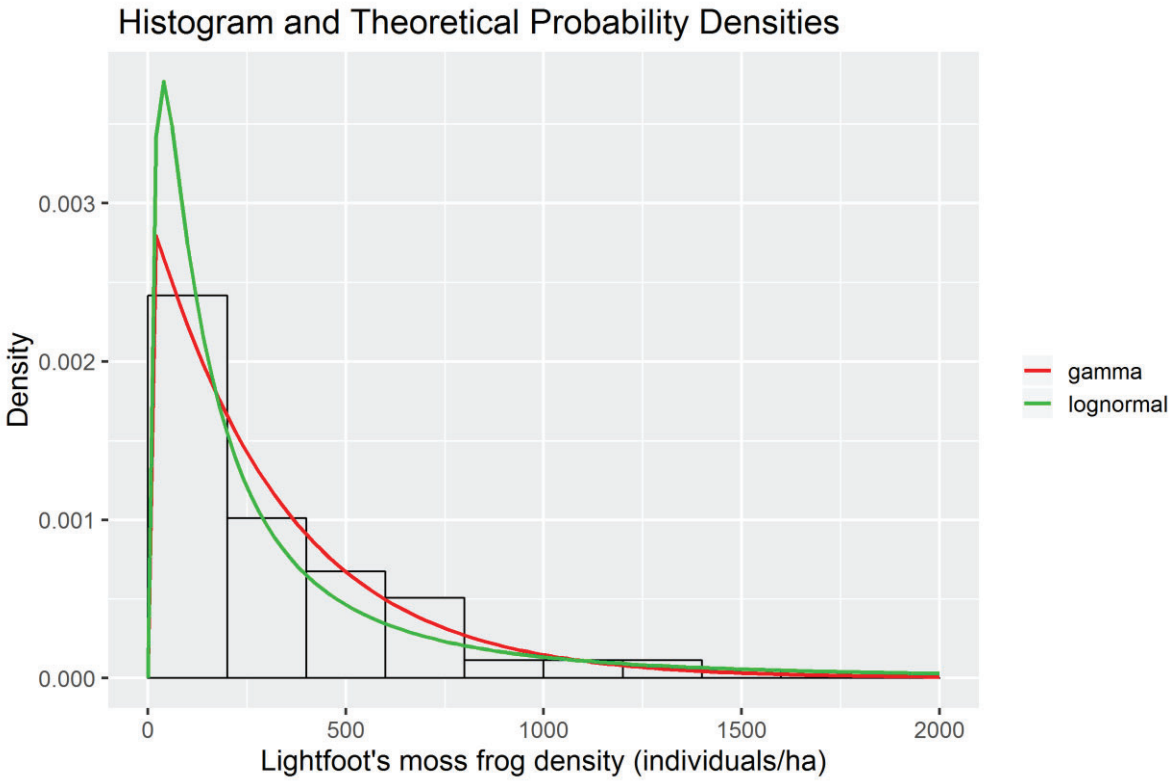


Figure 4.1 - Density plot of each fitted distribution with the histogram of population density data.

The gamma distribution was also selected by the AIC as the best performing distribution, with a difference in AIC scores of 18.32. Therefore, the latter was used to model the second component of the hurdle model. For each of the hypotheses formulated, two-part models were fitted. That is, the first part consisted of modelling of the species were present or not and the second part modelled the abundance of the species given its presence in terms of the explanatory variables expressed under the five hypotheses. I fitted five part model each to the presence/absence and density data and combined them to form 25 complete hurdle models by adding the AIC values from the two separate models (Table 4.1).

Table 4.1 - AIC values scored by the two separate components of the Hurdle model and the final AIC values of the Hurdle model. A, B, C, D and E are the five model Structures specified for each hypothesis. The hurdle model formed from climate variables 'A' for both presence/absence and density components is denoted as 'AA', while a presence/absence component using topographic variables 'B' combined with a density component using vegetation variables 'C' is denoted as 'BC'.

	Presence Model	A Climate	B Topographic	C Vegetation	D Wetland	E Null Model
Density Model	AIC	226.07	225.59	228.24	208.06	232.79
A Climate	69.683	295.75	295.27	297.92	277.74	302.47
B Topographic	69.854	295.92	295.44	298.09	277.91	302.64
C Vegetation	68.079	294.15	293.67	296.32	276.14	300.87
D Wetland	66.517	292.59	292.11	294.76	274.57	299.31
E Null Model	64.715	290.79	290.31	292.96	***272.78***	297.5

The hurdle model 'DE' with the null model fitted to the abundance data and the wetland model fitted to the presence data had the lowest AIC among the models. However, the AIC value for model 'DD' was not much higher, therefore the ΔAIC is computed, that is, the difference in AIC between each hurdle models and the best. We then also looked at the Akaike Weights, w_i (Anderson et al., 2001), which are interpreted as strength of evidence for a particular model to be the best, relative to other models formed (Burnham and Anderson, 2003). Those results are summarized in Table 4.2.

Table 4.2 - Results of AIC analysis of 25 competing models. Number of parameters including constant (K), -2 log likelihood scores (-2LL), AIC scores, differences among AIC scores- $\Delta_i AIC = [AIC_i - \min(AIC)]$, and AIC weights ($W_i$) – representing the relative likelihood of each model.

Model	-2LL	K	AIC	ΔAIC	W_i
AA	277.75	8	295.75	22.98	<0.001
AB	277.92	8	295.92	23.15	<0.001
AC	278.14	7	294.15	21.38	<0.001
AD	278.58	6	292.59	19.81	<0.001
AE	278.78	5	290.79	18.01	<0.001
BA	277.27	8	295.27	22.50	<0.001
BB	277.44	8	295.44	22.67	<0.001
BC	277.67	7	293.67	20.90	<0.001
BD	278.1	6	292.11	19.33	<0.001
BE	278.3	5	290.31	17.53	<0.001
CA	291.92	7	297.92	25.15	<0.001
CB	282.09	7	298.09	25.32	<0.001
CC	282.32	6	296.32	23.55	<0.001
CD	282.75	5	294.76	21.98	<0.001
CE	282.95	4	292.96	20.18	<0.001
DA	263.74	6	277.74	4.97	0.048
DB	263.91	6	277.91	5.14	0.044
DC	264.17	5	276.14	3.37	0.106
DD	264.57	4	274.57	1.80	0.232
DE	264.77	3	***272.78***	**0**	**0.571**
EA	290.47	5	302.47	29.70	<0.001
EB	290.64	5	302.64	29.87	<0.001
EC	290.86	4	300.87	28.10	<0.001
ED	291.30	3	299.31	26.54	<0.001
EE	291.5	2	297.5	24.73	<0.001

Looking at ΔAIC makes it clear that model 'DE' and 'DD' are close competitors for being the best model in the set. It is convenient to then look at the Akaike weights to have a better idea on the overall performance of the models over the whole set. Model 'DE' has 57.1% of the total weight and model 'DD' has 23.2%.

Of the five a priori hypotheses tested for A. lightfooti occurrence, the wetland model had the greatest support, with estimated coefficients suggesting that a $1m^2$ increase in wetland area increases the odds of finding the species by $\exp(0.004) \approx 1$. However, although occurrence of *A. lightfooti* was more likely at high levels of wetlands, density of A. lightfooti decreased with increasing wetland area (per m^2), suggesting that the factors driving those two processes might differ.

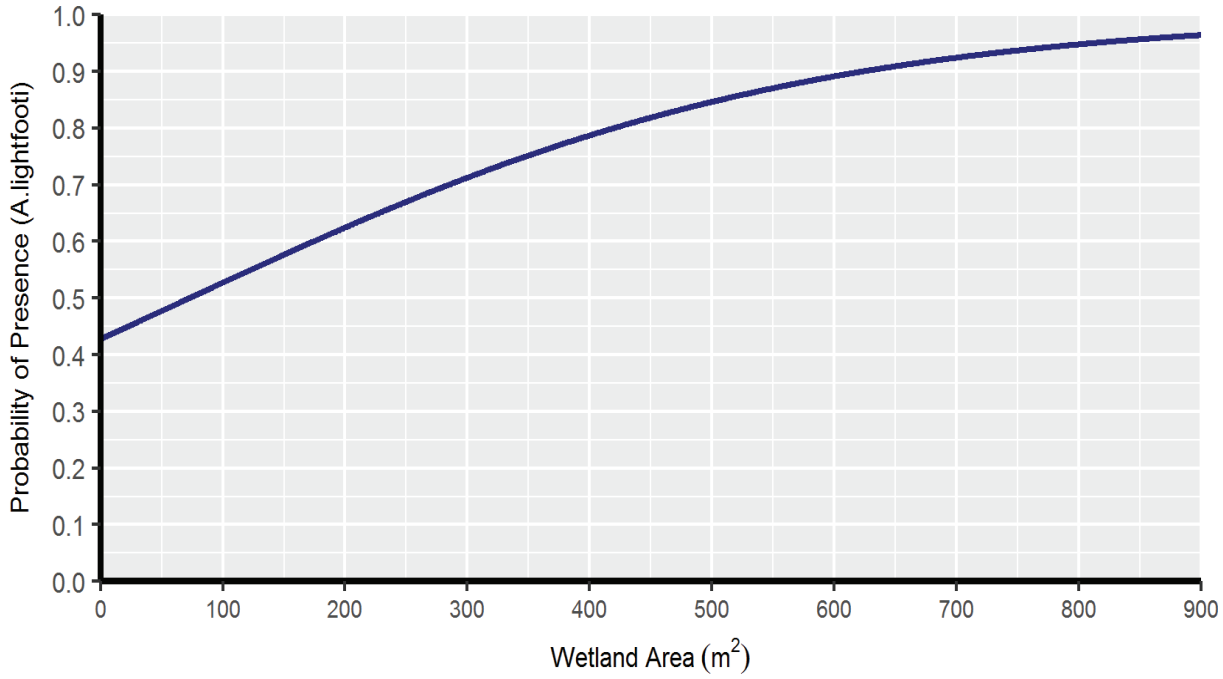
When investigating the topographic hypothesis, although this species is more likely to occur on east facing slopes, they are more inclined to occur at higher densities on steeper and west facing slopes. Similarly, under the vegetation model, the lightfoot's moss frog is less likely to occur under other type of vegetation but when found there they occur at higher densities; further suggesting that processes influencing occurrence and density were dissimilar.

There is clear evidence that of the five a priori hypotheses tested for A. lightfooti occurrence, the wetland model had the greatest support whereas the null model was the most parsimonious describing A. lightfooti density. Thus, the two model emerge as the best hurdle model and are combined to give the expected lightfoot's moss frog abundance, together with bootstrap-based confidence limits. Figure 3.2 shows how the results of the two models are combined. Using the bootstrap samples that were generated to calculate 95% CI, we estimated the bias in expected abundance of the lightfoot's moss frog for the range of wetland area shown in Figure 3.2.

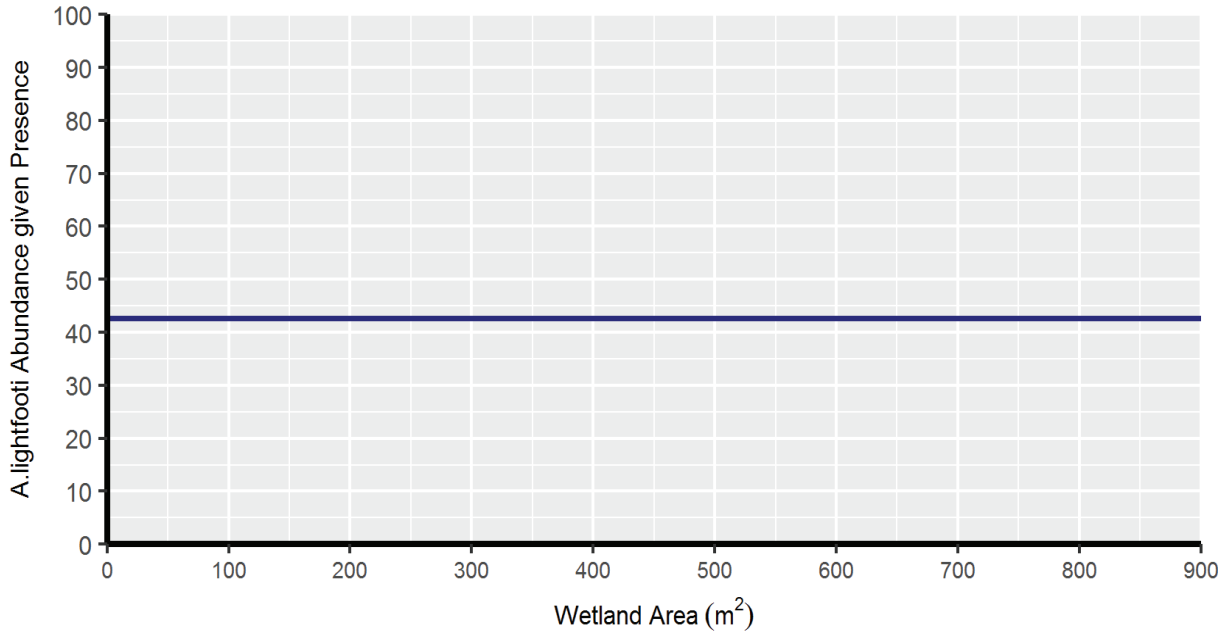
Table 4.3 - Estimates and standard errors of the coefficients for the explanatory variables of interest in the ten logistic and gamma regression models. A, B, C, D and E relates to the five model structures specified for each hypothesis.

		Presence model		Density model	
Model	**Parameters**	**Estimate**	**SE**	**Estimate**	**SE**
A	**Intercept**	2.611	2.955	5.558	1.184
	Winter Temp	-0.783	0.817	-0.835	0.331
	Winter Temp²	0.038	0.056	0.065	0.023
	Solar Rad	0.0004	0.0002	0.0002	0.0001
B	**Intercept**	0.738	0.274	3.699	0.115
	Eastwest	0.760	0.272	-0.371	0.128
	NorthSouth	0.119	0.286	0.304	0.129
	Slope	-2.062	0.840	0.511	0.544
C	**Intercept**	-1.386	0.791	2.962	1.186
	Vegtype 11	0.951	0.880	1.195	1.199
	Vegtype 16	1.746	0.811	0.636	1.190
D	**Intercept**	-0.289	0.179	3.867	0.101
	Wetland area	0.004	0.001	-0.0005	0.0003
E	**Intercept**	0.132	0.155	3.7530	0.081

a)


Probability of Presence (A.lightfooti)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
100
200
300
400
500
600
700
800
900
Wetland Area (m2)

b)


A.lightfooti Abundance given Presence
100
90
80
70
60
50
40
30
20
10
0
0
100
200
300
400
500
600
700
800
900
Wetland Area (m2)

c)

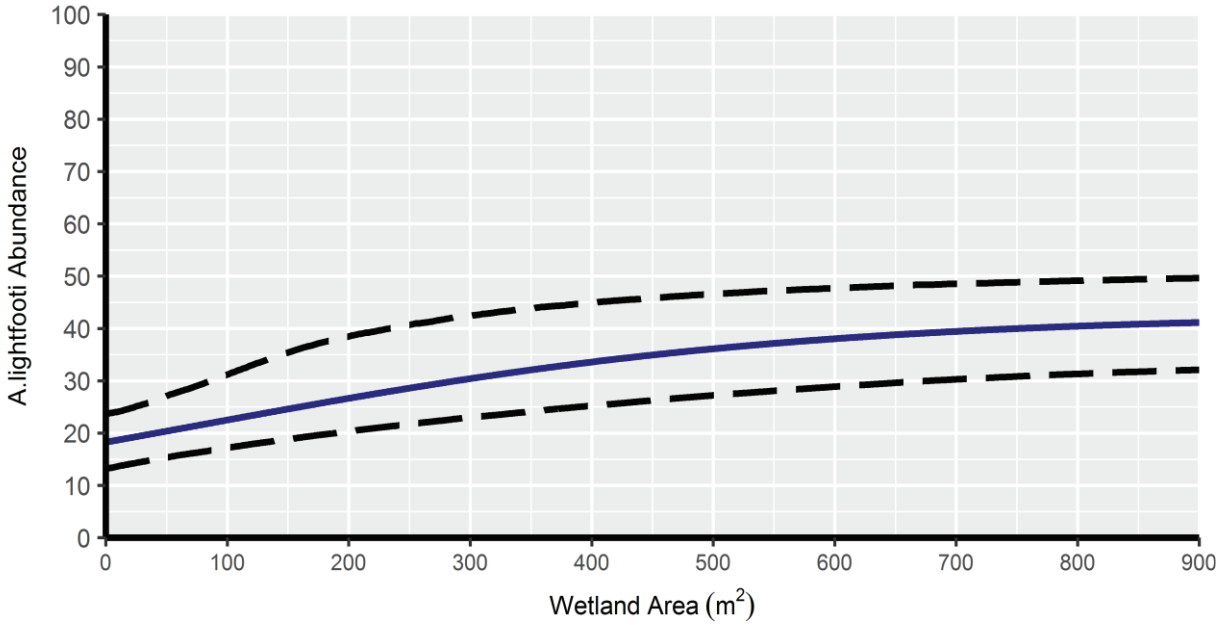


Figure 4.2 - Estimates of (a) probability of presence, (b) expected abundance given presence and (c) expected abundance of *A.lightfooti* (dashed lines are 95% confidence limits), plotted against area of wetland. The predictions are for wetland area ranging from 0 to 900 m^2 and thus abundance is measured in population density per m^2.

4.5 Discussion

It has been long recognized that the abundance and distribution of a species are closely related (Brown, 1984). With no quantitative information on the abundance of a species, ecological hypotheses on species abundance were drawn on mere assumptions about the species distribution (Sagarin et al., 2006). One such assumption is that the same environmental factors that determine species occurrence also determines abundance patterns inside the range (Brown, 1984; Gaston et al., 2000). In this chapter, I tested the ability of five model structures to describe occurrence and density of *A. lightfooti* across their entire range. The model structures represented a-priori hypotheses about the factors that potentially drive the frogs' distribution and the goal of this chapter was to examine whether the same factors drive occurrence and density of this species.

It was found that the variables explaining variation in occurrence were not always those explaining variation in density. The environmental conditions examined, although predictive of occurrence, were generally poor predictors of *A. lightfooti* across all ranges of density. Presence of the Lightfoot's moss frog was largely explained by topographic and wetland variables. In contrast, predictions of density were only weakly related to these same environmental factors and in some cases with opposite sign. Basically, it appeared that unmeasured factors were largely moulding observed patterns of *A. lightfooti* population density. Likewise in their study on bracken fern trees, Nielsen et al. (2005) found similar contradicting results which they attributed to unmeasured factors such as life and site history, that ultimately determine abundance, once occupancy was established.

For instance, even though its' habitat is largely protected, the spread of alien vegetation can change microhabitats used by *A. lightfooti* and could be a significant determinant of this species abundance. Clusella-Trullas and Garcia (2017) found that there are profound negative impacts of invasive alien species on ectotherms species diversity. Moreover, the recent drought that gripped South Africa would be expected to severely affect wetland capacity and longevity. Similarly, the time since the last intense fire would also contribute toward explaining the longevity of seepages, where the species has mostly been known to occur (Channing, 2001). Thus, it seems reasonable to assume site history could play a significant role in shaping *A. lightfooti* density.

However, it should be taken into account that the sites were not selected completely at random (see Section 2.2). According to the abundant-centre theory (Sagarin and Gaines, 2002) – abundance should be the highest where the environment is most favourable for the species, thus sampling was biased towards sites that were thought to be more suitable for this species. Thus, the null model represents an estimate of mean density for the sampled sites. However, with this study, we have learned that the abundant-centre theory might not hold in the case of *A. lightfooti* as the hypotheses examined in this chapter showed that

although the habitat is highly suitable at certain sites, there might not necessarily occur at high densities at those sites. However, one possible explanation is that – some sites visited during the survey period had low predicted habitat suitability and reported high densities.

One possible consideration is to look at the temporal variation in the relationship between distribution and density. For instance, reproduction in this species may relate weakly to the environmental conditions, but could be paramount in determining distribution and density during certain times (Chen et al., 2005). Moreover, as resources become patchier and limited, the strength of the relationship between density and distribution is usually expected to increase (Nielsen et al., 2005). On the other hand, if resources are freely available and of equal quality throughout a species range then there is no reason for them to amass in patches, which would make modelling species abundance difficult.

Ultimately, the idea of using hurdle models for positively skewed data that contain a large proportion of zeros is not new. Welsh et al. (1996) illustrated the concept with data for the abundance of Leadbeter's Possum. The approach allowed us to suitably model the presence/absence and density data using logistic and gamma regression respectively to estimate the two parts of the hurdle models. The approach also allowed us to use different sets of explanatory variables to model the two parts of the model, hereby leading to a better understanding of the processes determining the species distribution and abundance within its range.

Chapter 5 – Modelling the distribution of *A. lightfooti* based on opportunistically collected presence-only data

5.1 Introduction

In the previous chapter I investigate specific ecological hypotheses on the potential environmental drivers of the distribution and density of *A. lightfooti*. In this chapter, I model the distribution of this species based on opportunistically collected data (section 2.3) with the view to combine the resulting information with aSCR survey data (section 2.2) in the following chapter.

SDMs have become increasingly popular in the research fields of ecology and nature conservation planning (Maggini, 2011), whereby these models are used to simulate species' ranges from a limited set of known observations (Elith and Leathwick, 2009). However, efforts to parameterize SDMs have often created a dilemma between the quality and quantity of data available to fit models (Pacifici et al., 2017). A vast majority of species data that is available today consist of presence-only data collected under non-standardized designs (Phillips et al., 2009). There are numerous discussions regarding the way those data are analysed. One framework that has been developed by Pacifici et al. (2017) and used in this study - is to integrate the information from the opportunistically collected presence records as a constructed covariate with the higher quality data, in this case being the aSCR survey data (section 2.2).

Moreover, with the different statistical algorithms available for modelling, recent studies have shown that disparities among different model prediction can be large making model selection difficult (Guisan et al., 2017; Marmion et al., 2009). An alternative is to use an ensemble of models. The concept of ensemble modelling is to avoid selecting one single best model but instead to use a group of models for inference (Marmion et al., 2009). The different resulting 'habitat suitability

indices' are then averaged to get a single value per site (or 'grid cell', Araújo and New, 2007).

The rationale behind ensemble modelling is that different algorithms give different predictions and levels of accuracy under different circumstances and there is no single "best" model (Elith et al., 2006). In other words, the presence of a species might be well classified by some models and misclassified by others, thus making use of an ensemble model one can reduce the predictive uncertainty of a single model by combining predictions (Grenouillet et al., 2011; Marmion et al., 2009). To date, a substantial amount of studies have shown that ensemble techniques substantially improve the accuracy of species distribution predictions (1995; Araújo and New, 2007; Grenouillet et al., 2011; Marmion et al., 2009; Stohlgren et al., 2010).

Thus, the aim of this chapter was to produce a prediction map from opportunistically collected presence records of A. lightfooti using an ensemble of SDMs. The resulting habitat suitability map derived from this 'lesser quality' opportunistically collected data is then used as baseline information to integrate with the survey data of 'higher quality' in the following chapter.

5.2 The Data

Species Data

For the presence-only dataset, a total of 494 opportunistically collected occurrence records of the *A. lightfooti* was acquired dating from 1933 to 2015 (Figure 2.3). The data were collated from various sources (Appendix A). The records are spread latitudinally and longitudinally throughout the study area. Reported localities for a species that were not verified or treated as doubtful, were eliminated from the modelling process. For example, locations that were found on sea or in urbanized areas, where the species do not occur were removed. If several

records were found within the same 30 x 30 m^2 area were considered as duplicate and were also eliminated.

All location coordinates of *A.lightfooti* were checked against published ranges of the species (Turner and De Villiers, 2007). Knowledge of the habitats and regions greatly facilitated the assessment of sampling bias (Merow et al., 2013a). Experts with this knowledge were consulted when deciding on the final datasets to be used in the model. After data cleaning, a total of 366 records were kept for further analysis.

Environmental Data

The next step in formulating SDMs is to obtain appropriate spatial environmental data that describe a suitable composition of the *A.lightfooti* environmental requirement (Austin, 2007; Miller, 2010). Austin et al. (2006) investigated the performance of SDMs with regard to the choice of environmental predictors used and found that the latter had a major influence on SDM performance.

The type of environmental variables used in this study is species-specific and relies on expert's judgement. Environmental variables used to model species distribution based on presence only data are (see Section 3.3 for description):

- Elevation
- Slope
- Vegetation types
- Vegetation subtypes
- Vegetation communities
- Roughness
- North-South
- East-West
- Mean Min Winter Temperature
- Mean Max Summer Temperature
- Wetland types
- Summer Solar Radiation
- Winter Solar Radiation
- TPI (Topographic Position Index)

5.2 Ensemble Species Distribution Models Analysis

The biomod2 v3.3-7 package (Thuiller et al., 2016) in R statistical software (R Core Team, 2018) was used to fit the SDMs in this study. It was chosen due to the speed and ease of fitting multiple models available from the package itself and it provided a platform for ensemble forecasting of the species' distribution. It also provided quick graphical results of forecasts often needed for ecological interpretation of the distributions and model comparisons.

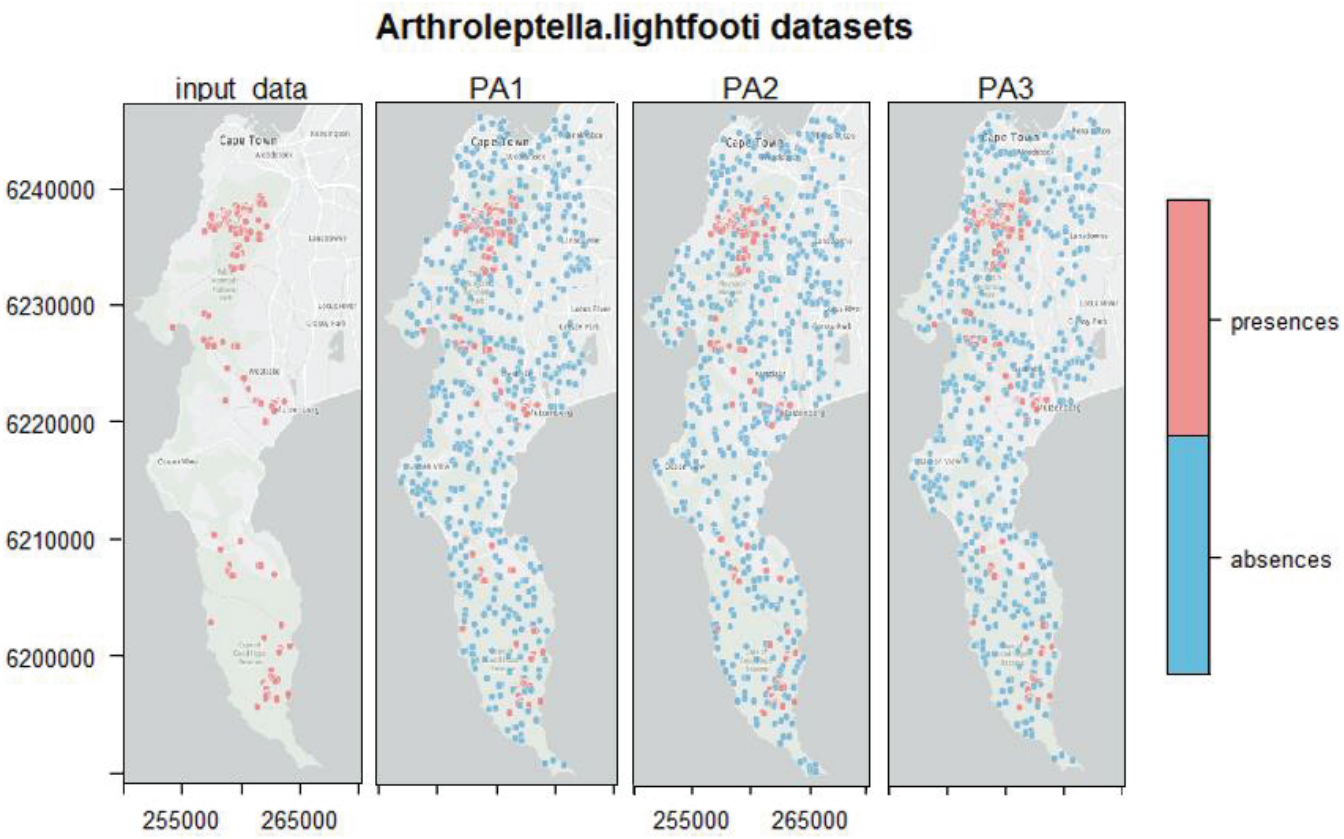


Figure 5.1 - Plot of species current distribution (red dots) and three sets of pseudo-absences (blue dots).

All models were calibrated with historic presence-only data and 1000 pseudo-absence points randomly selected from the entire area of the Cape Peninsula. Since this procedure implies a stochastic background point selection, it is recommended to build multiple sets of pseudo-absences data to prevent particular situations reflecting a specific sampling (Barbet-Massin et al., 2012) thus three different pseudo-absence dataset were created in the study (Figure 5.1). Each of the datasets has the species occurrence records together with an equal number of independently drawn pseudo-absences. Pseudo-absences were generated by randomly selecting points from all points within the study area excluding the available presence points.

The aim of this chapter was not to investigate the specific ecological interpretation of the environmental predictors, thus multicollinearity among predictors was not considered to be a problem (Hill and Judge, 1987).

All the models were fitted to the data using biomod2 (Thuiller et al., 2016). For each model, the following procedures were used:

A. Generalized Linear Models (GLM) (McCullagh and Nelder, 1989) with linear, quadratic and categorical terms but no interactions between covariates. A stepwise procedure was used to select the most significant variables using AIC (Thuiller et al., 2009).

B. Generalized Additive Models (GAM) (Hastie and Tibshirani, 1990) with a cubic spline smoother, which is a collection of polynomial of degree 3. Similarly, to GLM, an automated stepwise process was used to select the most significant variables (Thuiller et al., 2009).

C. Classification Tree Analysis (CTA) (Breiman, 2017) using the *rpart* library (Therneau et al., 2010). This procedure runs a 10 fold cross-validation to select the best trade-off between a high decrease of deviance and the smallest number of leaves (Thuiller et al., 2009).

D. Random Forest (RF) (Breiman, 2001) classification tree was fitted using 500 trees.

E. Generalized Boosting Model (GBM) (Ridgeway, 2007) with an interaction depth of 4, a learning rate of 0.001 and a total of 2500 trees respectively with five-fold cross validation.

F. ANN (Abraham, 2005) where BIOMOD uses the library *mnet* (Ripley et al., 2016). ANN was parameterized with the number of units in the hidden layer and weight decay optimized by cross-validation on model AUC. 5 folds cross-validation was used to find the both parameters.

G. MaxEnt (Phillips et al., 2004) model was calibrated using linear, quadratic, product, threshold and hinge features, with a regularization parameter of -1.50 and a 0.5 probability of presence in any cell within the study area.

In the absence of an independent dataset, a split-sample cross-validation approach (Guisan et al., 2017) was used for each model's evaluation. For validation, 20% of both the presence and pseudo-absence datasets were partitioned and set aside for cross-validation while 80% was used to calibrate the models. The split-sample procedure was then repeated three times over the three different pseudo-absence datasets.

Do different modelling methods select different types of environmental variables? To compare the importance of the explanatory variables across the models, a permutation procedure was implemented to examine the importance of the variables in the model. The technique consists of making standard predictions from the calibrated models and then the variable under investigation is randomly permutated and a new prediction is made (Thuiller et al., 2009). The correlation score between the standard predictions and the new predictions give an estimation of the variable's importance (Naimi and Araújo, 2016). Here this procedure is repeated three times for each variable independently. A good correlation score

indicates that the contribution of a variable to the model is low (Ishwaran, 2007; Naimi and Araújo, 2016). The importance scores for all predictors were standardized by calculating the proportion of each predictor's importance relative to the sum of all predictors' scores in each model, so that the total variable importance was equal to one for comparison purposes.

Each single model was run on the training partition and evaluated on the test partition by the commonly used, area under the receiver of characteristics curve (AUC) in order to assess the predictive performance of the different modelling techniques (Hanley and McNeil, 1982). Each of these replicate models was projected on to the study area and the final projection for each modelling algorithm was achieved by averaging the nine replicate models across the cross-validation runs and pseudo-absences datasets.

A total of 63 (7 models * 3 PA-dataset * 3 cross-validation runs) algorithms were combined by taking the average prediction from each algorithm to build the ensemble model. The ensemble model was then evaluated over the entire dataset, which is the union of the three PA datasets. The single models and ensemble model were then compared using the AUC value.

5.4 Results

5.4.1 Ensemble Maps

The habitat suitability map for *A. lightfooti* (Figure 5.2) is a result of the averaged predictions across the seven statistical modelling techniques. The uncertainty of the resulting ensembles was also assessed by calculating and mapping the standard deviation between the predictions of each of the seven models (across the pseudo-absence datasets and cross-validation runs) composing the ensemble. The ensemble model performed well with an AUC value of 0.977.

The geographical areas with the highest habitat suitability are mostly on the top of Table Mountain National Park, Noordhoek region and along the south of the Cape Peninsula (Figure 5.3). A high variability can also be observed between the statistical methods in small areas across the edge of observed distribution.

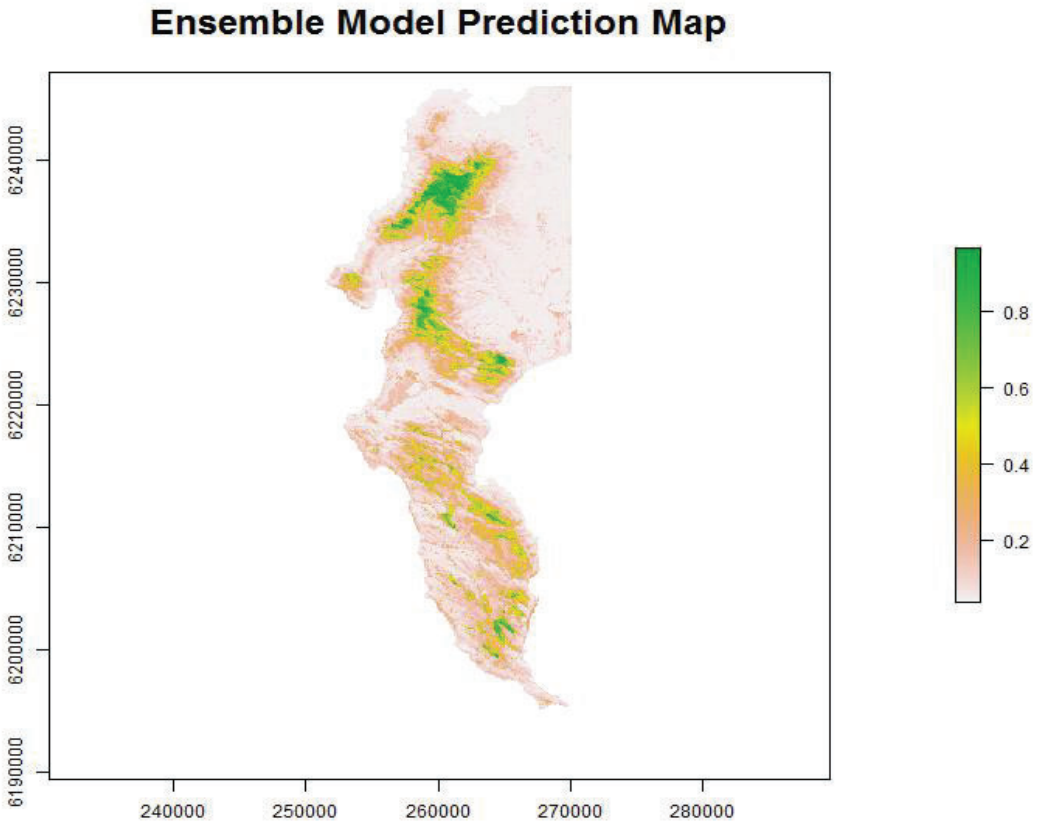


Figure 5.2 - Ensemble Model showing the predicted habitat suitability of the *A. lightfooti* obtained by averaging single models' predictions.

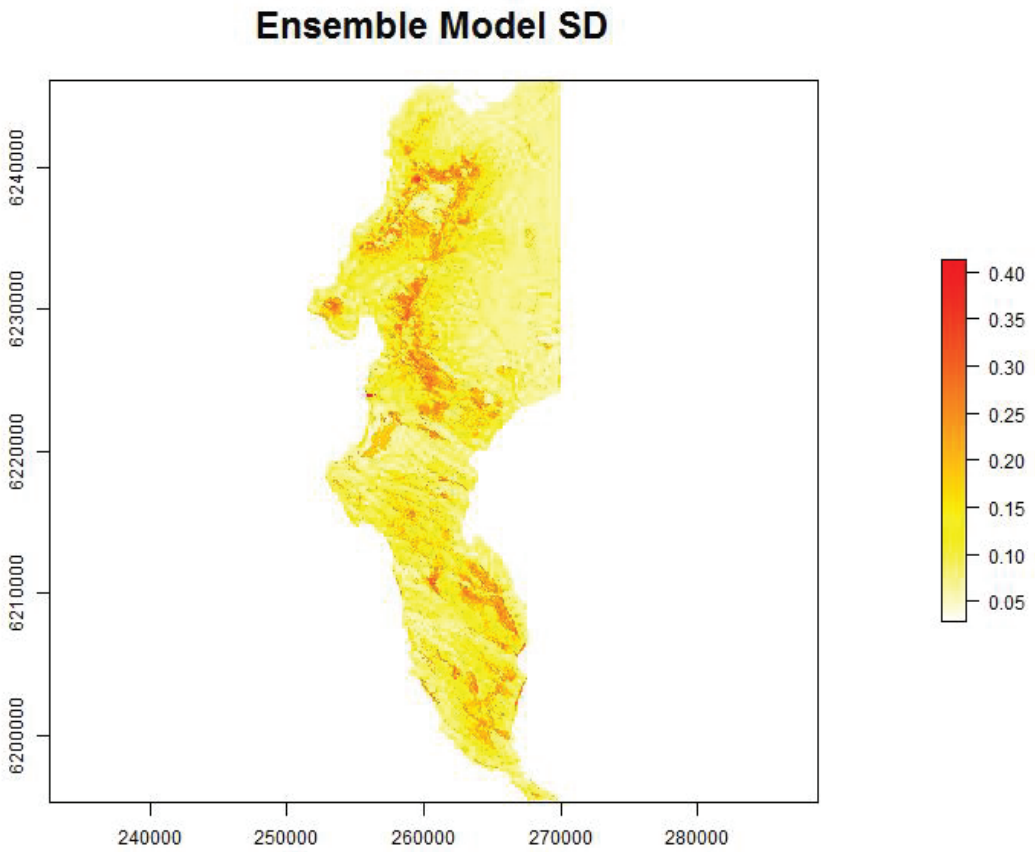


Figure 5.3 - The ensemble modelling uncertainty (SD). Red high (0.35 - 0.40) and yellow low (0.10 - 0.05).

5.4.2 Evaluation of Results across Modelling Algorithms

In this section, the individual algorithms are evaluated briefly in three different ways. Firstly, comparing the AUC values to assess differences in the general model fit. Secondly, comparing the geographical consistency of the maps produced by each of the algorithms to assess the predictions conformity across algorithms. Thirdly, comparing the contribution of the various environmental variables to the different models.

Overall, the seven single-SDMs showed good ability to predict observed distributions, with AUC values ranging from 0.768 to 0.971 (Appendix A). Figure 5.4 highlights the variability among the cross-validations runs and pseudo-absence datasets for each model. The AUC varies about 0.05 for ANN, which in turn reveals the largest variance compared to the other models. For models' evaluation

across the different cross-validation runs and pseudo-absence datasets, see Appendix A. CTA and GAM show very similar mean AUC and similar variances, while GBM and RF show the best AUC values and generally a smaller variance between runs and pseudo-absence datasets. The relative ranking of SDMs according to their AUC values showed that RF more frequently yielded the models with the highest predictive performance.

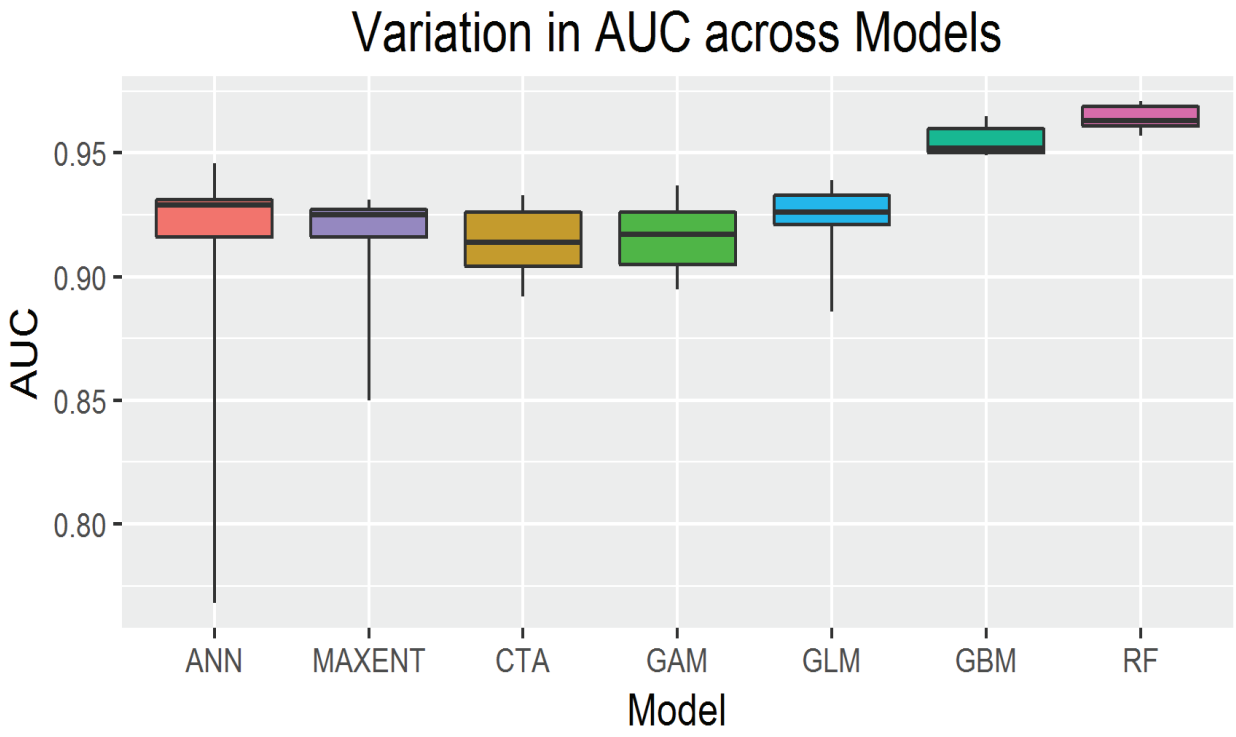


Figure 5.4 - Variation in the area under the receiver-operating characteristic curve (AUC) among the cross-validation runs and the pseudo-absence datasets between the different models.

Maps were produced to illustrate the geographical areas of species occurrence predicted by the different modelling techniques across the different cross-validation runs and pseudo absence datasets. The maps showed that the single-SDMs models made predictions that were broadly consistent with each other (Figure 5.5). The colours show the habitat suitability index, green high (0.8 to 1.0) and light pink low (0.3-0). All non-zero predictions are within the environmental range of the training data (i.e. the model is not predicting to novel environments).

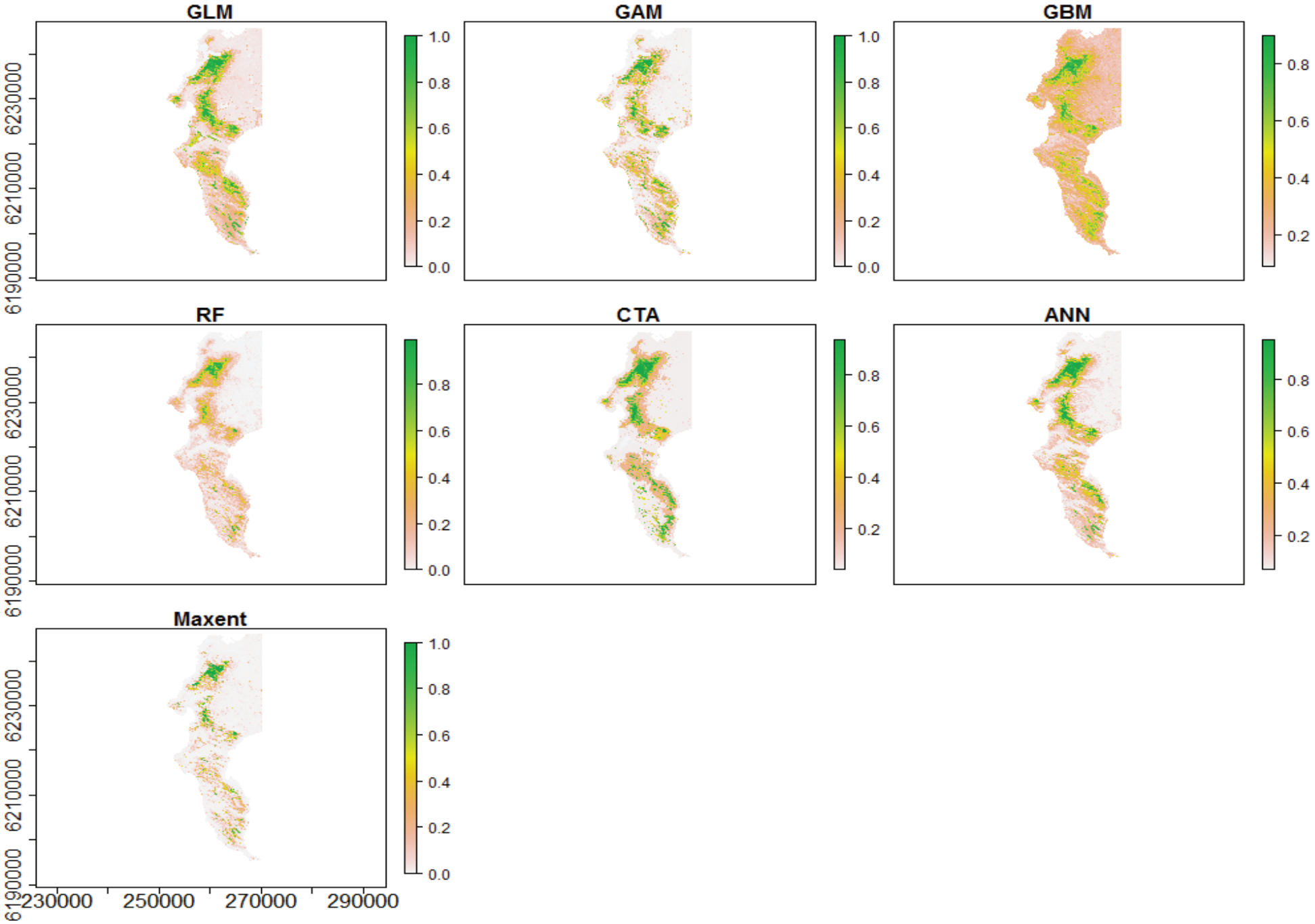


Figure 5.5 - Habitat suitability of the A. lightfooti predicted by the seven single-SDMs averaged across the different cross-validation runs and pseudo-absence datasets.

Moreover, we looked at the relative influence of the environmental variables in single models (Figure 5.6). In biomod2 the variable importance is measured as '1-correlation', thus the higher the score, the more important the variable. The variable importance was averaged over the different pseudo-absences datasets and cross-validations runs. A table with the average variable importance across the different pseudo-absences sets and cross-validation runs can be found in Appendix A.

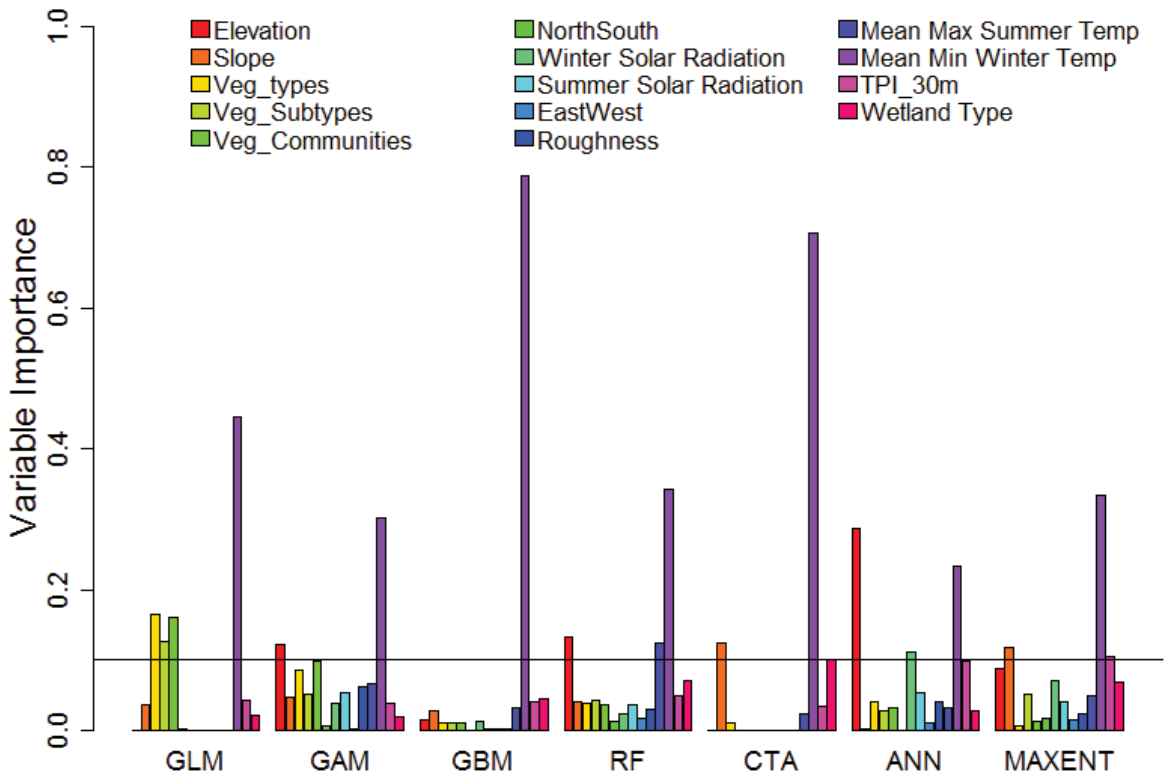


Figure 5.6 - Relative importance of the environmental variables used to predict the potential distribution of the *A. lightfooti*. Continuous black line refers to the mean relative importance value of the predictors across the seven models.

There was substantial variation in the importance of different environmental predictors among statistical models except for the mean minimum winter temperature variable which appears to be the most influential for all models except for ANN model. The elevation variable seems to be the second most influential. Across all seven models, only two to four variables out of 14 have values of relative importance higher than the mean importance value.

5.5 Discussions

The aim of this chapter was to produce a prediction map from opportunistically collected presence records of the *A. lightfooti*, such that this information can be combined with the survey data in the next chapter. I produced the prediction map using an ensemble of SDMs and my findings confirmed that ensemble forecast offers an improvement on any individual model (Marmion et al., 2009; Grenouillet et al., 2011; Thuiller, 2004).

The ensemble prediction map was built to account for projection variability among the single SDMs. The ensemble modelling uncertainty map shows interesting patterns and highlights where the individual models show divergent results. Interestingly, predictions where the model algorithms differ most in their predictions are mostly at the edges of the observed distribution. The unsettled nature of the independent models output justifies the combination method of prediction that produces an ensemble forecast. Marmion et al. (2009) and Grenouillet et al. (2011) also reached to similar conclusions

Based on AIC scores, my results showed that tree-based methods, such as RF and GBM, showed predictive abilities close to the ensemble model and this is consistent with earlier studies (Cutler et al., 2007; Marmion et al., 2009). RF and GBM are also based on the idea of ensemble modelling whereby hundreds of trees are built and the mean prediction is computed. Overall the AIC scores indicated fairly good discriminatory ability of all the individual models with most AUC values being greater than 0.8.

The predictor variables in this study were included so they cover suitable range of the *A.lightfooti* environmental requirement. The 'mean minimum winter temperature' variable appeared as the most significant variable across all seven modelling techniques. The second most influential variable was 'elevation', often selected by different models. One noticeable distinction is that CTA led to the

simplest single model, with only seven variables whereas ANN, Maxent, GAM and RF produced the most complex models.

An improvement for ensemble models worth considering is individual model formulation. Although an ensemble model is a good way of dealing with extreme variations in predictions across methods, the need to select explanatory variables and examining the necessity for polynomial or interaction terms (Li and Wang, 2013; Elith, 2015) is crucial. Nevertheless, occurrence records used in the study, collected in a non-standard way, produced a relatively realistic map as consulted with experts. In the next section, the ensemble forecast map will be used as an additional predictor variable to produce an updated map with more recent survey data.

Chapter 6 – Model-based density estimates of *A. lightfooti* across the Cape Peninsula using all available data

6.1 Introduction

The spatial prediction of abundance has been recognised as an important component of conservation planning (Franklin, 2010; Guisan and Zimmermann, 2000). Species abundance models have become a key tool for ecologists and conservation biologists (Elith and Leathwick, 2009). Today the greater availability of occurrence and abundance data has spurred on the use of species distribution and abundance models to make spatial predictions of abundance.

However, species distribution and abundance models can only perform as well as the data they are fitted to and are thus subject to biases in the underlying data (Merow et al., 2013a). Pacifici et al. (2017) suggested that these distribution and abundance models should be parameterized using high quality data. Data collected under standardized sampling designs that include a randomized and consistent sampling method which accounts for observer effort and detection uncertainty.

Huge amounts of data on many species are collected by citizen science programs (Giraud et al., 2016; Pacifici et al., 2017). Examples, include specimen collection data from museums and herbaria, and atlas records maintained by government agencies and non-government organizations. Unfortunately, these data are usually collected opportunistically and carry three main challenges – spatial bias in sampling intensity whereby some sites are more likely to be visited than others, false negatives[1] and false positives[2] (Giraud et al., 2016). A key challenge facing researchers today is the need to assemble and work with these low quality data from different sources (Pacifici et al., 2017).

[1] Non-detection of species at sites where they occur

[2] Species are wrongly reported at sites where they do not occur

However, building models using these data often violates key assumptions of the estimation procedures used to fit SDMs (Yackulic et al., 2013) which creates a tension between quality and quantity of data available to fit models (Pacifici et al., 2017). It is not uncommon now for projects to integrate any combination of data types into a single analysis. Dorazio (2014), Fithian et al. (2015) and Giraud et al. (2016) proposed the integration of presence-only or opportunistically collected data with standardized survey data. Fithian et al. (2015) considered a probabilistic model that allows for joint analysis of presence-only and survey data to exploit their complementary strengths. Dorazio (2014) developed a more general framework that accounts for detection errors and survey biases from the presence-only data.

This chapter aims at producing the first Peninsula wide population-density surface for *A. lightfooti* using all available data on the species, together with an uncertainty map. More specifically, this chapter aims at looking at whether the opportunistically collected presence-only records used to build a habitat suitability map of *A. lightfooti* in chapter can be used as a constructed covariate to a hurdle model to produce the best possible population-density surface. Compared to the preceding Chapter 4, where the aim was to examine whether the occurrence and density of the *A. lightfooti* were influenced by the same environmental factors based on specific *a-priori* hypotheses, this chapter uses a hurdle model that best fits the high-quality density data.

6.2 The Data

Our purpose is to estimate the relative abundance of the species at the different sites. The estimation is based on two data types, opportunistically collected presence-only records and systematically collected survey data. The two main problems with analysing opportunistically collected species records are not accounting for the observation process and spatial bias (Yoccoz et al., 2001). In this case, the density data were collected under design-based sampling protocols

and accounting for the observation process which overcomes those problems. The procedure thus covered the entire species' range without neglecting those areas which are rarely visited. The standardized dataset gives the population densities of *A. lightfooti* at randomly sampled sites of 900 metre squared, together with locations where no frogs have been observed. A total of 89 population density estimates was acquired, together with 78 recorded absences. A more detailed description of the dataset can be found under Section 2.2. In this study, this is considered as the 'high' quality data.

A second dataset consists of an opportunistic dataset, considered as the lower quality data here, characterized by a completely unknown sampling effort. It consists of 366 presence-only records from multiple sources obtained from 1933 to 2015. In Chapter 5, a habitat suitability map was constructed from these opportunistically collected data using an ensemble of species distribution models. A more detailed description of the dataset and the resulting habitat suitability map can be found under Section 2.3 and 5.4.1, respectively. The resulting habitat suitability map is used here as a constructed covariate. Moreover, other than being of lower quality, those opportunistically collected data were also used by Louw (2018) to stratify the sampling in order to focus the main effort to areas where the frogs were most likely to be present (see section 2.2). Thus, to account for this stratification, the opportunistically collected data should be used as a constructed covariate in the form of a habitat suitability map.

6.3 Analysis

The analysis proceeded in several stages.

1. Constructing a covariate model from presence-only records

A first stage analysis was carried out on the 366 opportunistically collected presence records obtained from 1933 to 2015, which resulted in a habitat suitability index map of *A. lightfooti*. A detailed description of the procedure and resulting map can be found in Chapter 5 of this thesis. The resulting map is

thereafter used here as a constructed covariate in this investigation. Therefore, a total of 16 predictor variables were used in the study; 15 environmental variables listed under section 2.4 and the constructed habitat suitability index map.

2. Fitting a hurdle model to the standardized abundance dataset

To examine whether the integration of a habitat suitability index map returns more accurate density estimates of the *A. lightfooti*, this stage involved building a base hurdle model to which the habitat suitability map would be incorporated as a covariate at a later stage. The base hurdle model was constructed following the same procedure described under Section 4.4, the standardized dataset is divided into two groups: one indicating whether the *A. lightfooti* was present or not at each site, the other showing the population density estimates together with the associated standard of errors for those sites where the frog was present. Similarly, both the presence data and abundance data were then modelled using logistic and gamma regression, respectively, but only in terms of the environmental variables.

Models containing main effects, all possible two-way interaction and quadratic terms were considered. For both types of model, a stepwise selection procedure was adopted, using the `glmulti` R package (Calcagno and de Mazancourt, 2010). The stepwise selection procedure is based on the AIC, where the variables are ranked according to their contribution to reducing the total AIC and retains the most parsimonious combination of variables (Franklin, 2010).

More specifically, starting with a forward stepwise regression, each environmental variable is fitted to the data independently and the model with the smallest AIC is retained. Then, each of the remaining variables is added one at a time and is retained only if they further reduce the AIC. The process perseveres until adding variables no longer reduce the AIC. Furthermore, among the selected variables, all possible interaction and quadratic effects are tested and those further reducing the AIC are kept in the model. Adequacy of the regression models were examined

by inspection of the residuals. In both cases, models with the lowest AIC scores were retained for further analysis.

3. Data Integration

The hurdle model is a two-part model, which allows different sets of explanatory variables to be used to build the two different parts of the model. Thus, four model structures were constructed: (1) a base model – where the habitat suitability index map was added in neither part of the hurdle model, (2) where the habitat suitability index map was added solely to the 'presence/absence' part of the hurdle model, (3) where the habitat suitability index map was added solely to the 'density' part of the hurdle model and (4) where the habitat suitability index map was added to both part of the hurdle model.

4. Model Evaluation

To compare the predictive performance of the hurdle model, a leave-one-out cross-validation procedure was applied. The parameter estimation is performed on $(n-1)$ of the n observations and then the performance of the fitted model is tested by predicting the n^{th} observation and by computing the squared error. The squared error is computed by deducting the predicted observation from the actual observation and squaring the result. So, in this procedure, the n^{th} observation is the test set and the other $(n-1)$ observations are the training data for optimising the parameters of the algorithm. Then repeating the process n^{th} times, each time leaving out a different observation to use as the single test case. In this case, $n =$ 167 (89+78) and thus the process is repeated 167 times and each time the squared error is measured.

As a final measurement of the quality of the model, the mean squared error (MSE) is computed. The mean squared error measures the expected squared difference between the predicted value from the model and what is truly observed. The predicted mean square error (MSE) is then cumulated after each cross-

validation iteration. The process is then repeated for the four hurdle models and the cumulated MSE are compared across the models. The model with the lowest cumulated MSE is then regarded as the best predictive model.

5. Mapping *A. lightfooti* Population Density

Finally using the best performing hurdle model, the expected *A. lightfooti* population density estimates are predicted across the study area, excluding sites in urbanized areas where the frogs do not occur. A map indicating projected population density should be accompanied by some measure of uncertainty (Guisan and Zimmermann, 2000). Thus, using non-parametric bootstrapping, standard deviations for the estimate of expected population density using the procedure described under Section 4.4 is constructed.

The model was required to predict to places not sampled in the training data, motivating the need for a measure of environmental similarity between new environments and those used in the training dataset. A Multivariate Environmental Similarity Surface (MESS) map, proposed by Elith et al. (2010), was used to identify and visualize novel environmental conditions. The method calculates the similarity of any given point in the region of projection to a reference set of points (e.g. points used to train the model) with respect to the chosen predictors (Elith et al., 2010). In other words, taking a hyper-dimensional box-like viewpoint, the method analyses the environmental coverage of one predictor variable at a time and reports how novel those conditions are outside the given defined covariate space (Zurell et al., 2012). MESS maps present the user with a quantitative measure of projection uncertainty by helping to identify extrapolated areas (Mesgaran et al., 2014).

6.4 Results

Following the stepwise regression procedure to find a model that fits the presence/absence and density data well, the retained logistic regression model included two main effect terms: Mean minimum winter temperature and wetland area, and a quadratic term for wetland area. The model scored an AIC value of 190.6. For the population density data, the selected gamma regression model proved to be the null model with an AIC value of 95.7. These two components were thus combined to form the base hurdle model to estimate the expected density of *A. lightfooti*.

Table 6.1 - The best candidate models considered for *A. lightfooti* occurrence and density modelling. Model number and structure are provided. In the table Winter Temp refers to Mean Minimum Winter Temperature.

Model Number	Model Structure	AIC
Logistic Regression Models		
1.	Winter Temp + Wetland area + (Wetland area)^2	**190.57**
2.	Eastwest + Winter Temp + Wetland area +(Wetland area)^2 + Eastwest * Winter Temp	190.71
3.	Eastwest + Winter Temp + Wetland area +(Wetland area)^2	191.31
4.	Winter Temp + Wetland area + (Wetland area)^2 + Winter Temperature*Wetland area	192.54
5.	Eastwest + Winter Temp + Wetland area +(Wetland area)^2 + Eastwest * Wetland.area	193.29
6.	Eastwest + Winter Temp + Wetland area	202.39

Gamma Regression Models		
1.	'Null'	**64.72**
2.	Winter Temp	66.35
3.	Wetland area	66.52
4.	Winter Temp + (Winter Temp)^2	67.84
5.	Wetland area +(Wetland area)^2	68.36
6.	Winter Temp + Wetland area +(Wetland area)^2	70.03

Results from the cross-validation procedure showed that all three hurdle models that incorporated the habitat suitability index map as a covariate had larger MSE than the hurdle model excluding the map as covariate (Table 6.1). Although Stone (1977) showed that the leave-one-out cross-validation and the AIC are asymptotically equivalent, the AIC in itself cannot be interpreted as the predictive accuracy of a model but is useful when comparing models. However, the MSE can be interpreted as a measure of the distance between predicted and actual observations and is therefore used in this chapter to assess the predictive accuracy of the models. The smaller the MSE the closer the model's prediction is to the actual data, indicating that adding the habitat suitability index map increases the mean squared error in predictions. The different map resulting from the other three hurdle model can be found in Appendix A.

How can habitat suitability not be a good predictor? In view of those results (Table 6.2), further investigation revealed a poor correlation between the observed density estimates and the habitat suitability indices (Spearman's Rho correlation of 0.266). Although the habitat suitability map was uninformative on the density

part of the hurdle model, one would expect that the suitability index at least predicts occurrence better than density since the ensemble SDMs were fitted to presence-only data. Thus, the difference between the predicted presence probabilities from the presence/absence part of the hurdle model and the habitat suitability was mapped (Appendix A). Areas where the two maps mostly disagree correspond to areas where the species were historically present (from the opportunistically collected data) but recorded as absent during the 2016 and 2017 detailed survey and vice versa. That is, areas where they were not historically observed (or areas that were not sampled in the opportunistically collected data), the more recent survey recorded them as being present.

Using the best performing hurdle model; the model that did not incorporate the habitat suitability index map as a covariate was used to predict the expected density of A. lightfooti across the study area (Figure 6.1 – Left panel). The map shows *A. lightfooti* density per 900m^2 grid cell. The highest concentration of the expected species population was found to be mostly on the top of Table Mountain, Noordhoek, Silvermine and across the south east region of the Cape Peninsula. Estimates were generally lower towards the coastal areas of the study area. Expected density of the *A. lightfooti* ranged between 5.5 and 42.2 per 900m^2, with a total expected population of 3 591 634 individuals across its entire range (95% CI: 1 988,763; 5 476 754). The map shows 'spikes' in density estimates due to the wetland surface that change rapidly across the Cape Peninsula. For example, one grid cell of 30m^2 can contain some wetland surface while the grid cell next to it can be free of any wetland surface.

Table 6.2 - Mean Squared Error (MSE) estimates of expected population density for the *A. lightfooti* across different covariate models. Together with the number of parameters including the constant (K), -2 log likelihood scores (-2LL) and AIC scores.

	Hurdle Model Structure		K	-2LL	AIC	MSE
	Logistic Regression	Gamma Regression				
1	Winter_Temp + Wetland area + (Wetland area^2)		5	243.28	255.26	**105 180.5**
2	Winter_Temp + Wetland area + (Wetland area^2) + Habitat Suitability Index map		6	243.10	257.10	105 211.4
3	Winter_Temp + Wetland area + (Wetland area^2)	Habitat Suitability Index map	6	243.05	257.06	105 249.7
4	Winter_Temp + Wetland area + (Wetland area^2) + Habitat Suitability Index map	Habitat Suitability Index map	7	242.86	258.87	105 305

Prediction uncertainty map (Figure 6.1 – Right panel) is the standard deviation of predicted density estimates from 1000 bootstrapped sample. Areas of high uncertainty (large confidence interval) were mostly observed on the southern part of the Cape Peninsula. As demonstrated by the mean minimum winter temperature MESS map (Figure 6.2 – Left panel); areas of high dissimilarity to

the surveyed points were observed on the southern part of the Cape Peninsula. Areas of moderate prediction uncertainty are the areas where the winter temperature MESS map (Figure 6.2 – Left panel) indicated the greatest dissimilarity. The most similar environments were observed mostly on the top of Table Mountain and other mountains of the Cape Peninsula. The wetland area Mess map (Figure 6.2 – right panel) shows high similarity areas but also small areas showing no similarity. Looking at the prediction uncertainty map (Fig 5.1 – left panel) we can observe that the highest uncertainty areas are associated with those small areas.

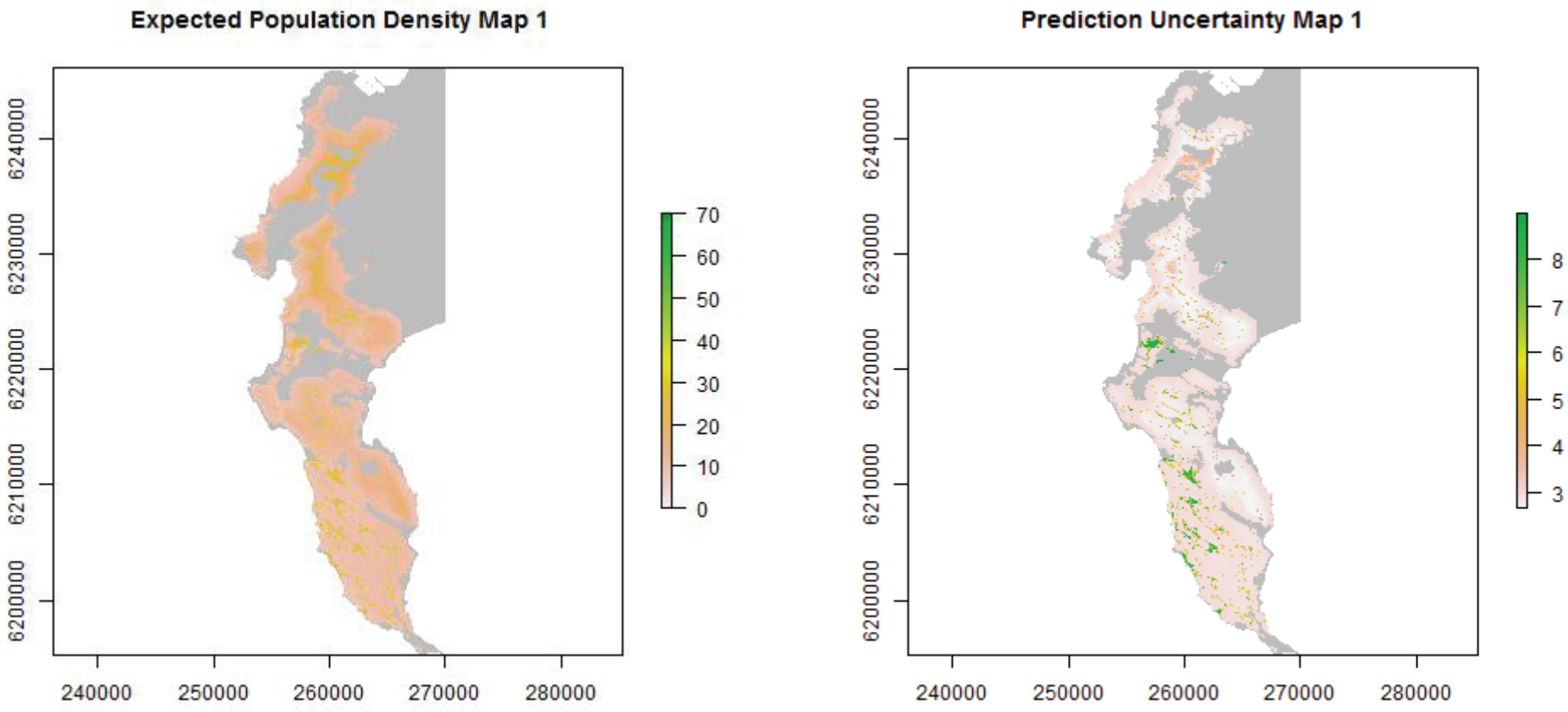


Figure 6.1 - Current expected population density estimates of the A. lightfooti across the Cape Peninsula (right panel), together prediction uncertainty. Green areas to low pink areas indicates high to lower estimates. The grey areas represent urbanized lowland areas where the frog does not occur.

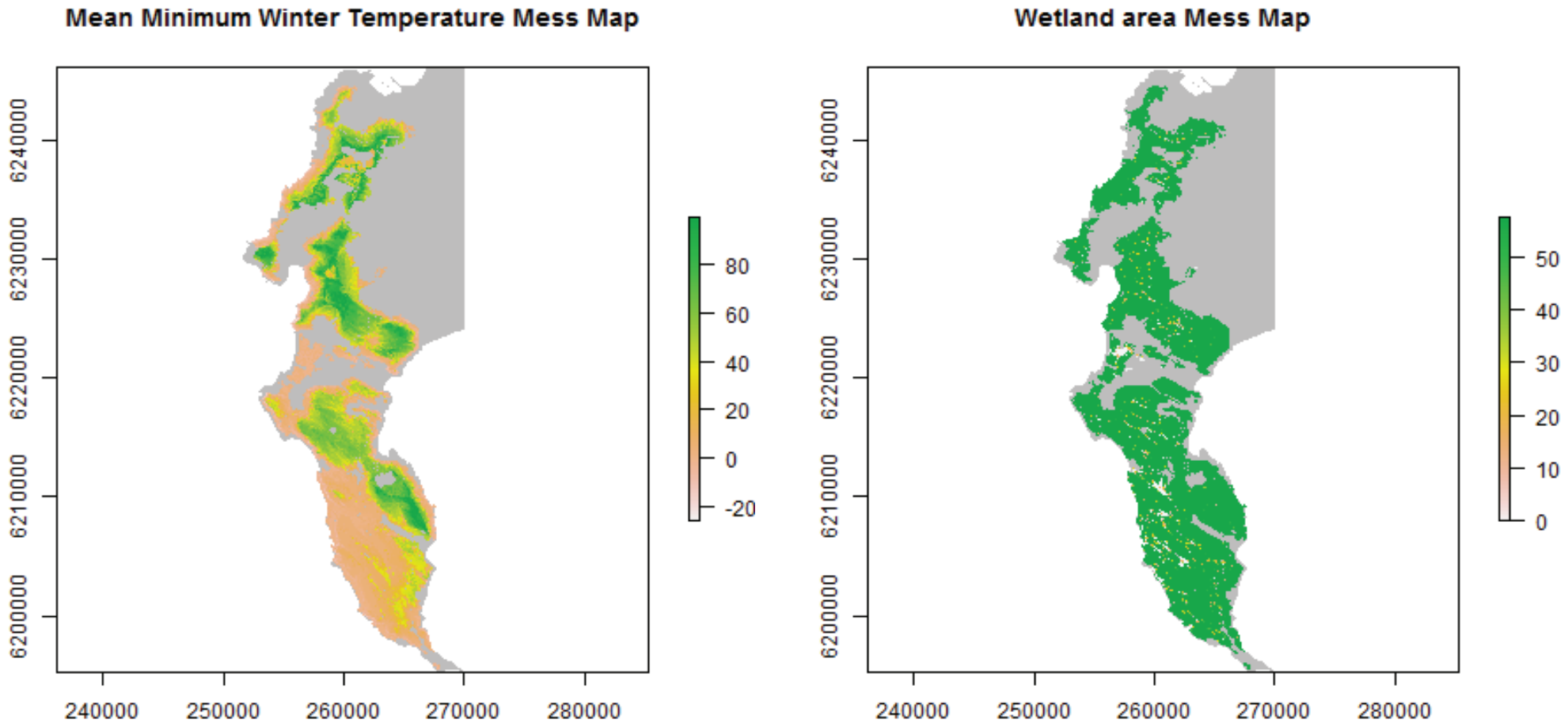


Figure 6.2 - Multivariate Environmental Similarity Surface (MESS) map for Mean Minimum Winter Temperature (left panel) and Wetland area (right panel) variables. Scale on MESS map: positive values shows analogous area with respect to each variable; negative values correlate to dissimilar environmental areas (pixels) from the surveyed points and areas around zero represent area where the variable is on the edge of the training range of the surveyed points.

6.4 Discussion

This study provided the first expected population density estimates of the Lightfoot' moss frog across the whole Cape Peninsula with a global population size of approximately 3.5 million individuals. The confidence interval around this estimates ranged from ~ 2 million and ~5.4 million individuals. However, this estimate derives from a model-based approach where the data were collected under a stratified sampling design (see section 2.2) whereby the main sampling effort was assigned to sites where habitats were most suitable.

Sampling locations for the density dataset were initially selected based on a stratified random sampling design to allow for a design-based inference of the abundance of *A. lightfooti* across its range. The design-based approach uses the probability of inclusion for each site that originates from the sampling design (Shi et al., 2016). However, in this case, if no frogs were heard at the target site, the next closest location of calling frogs were located by walking in circle around the target site (see section 2.2). The mean density empirically observed during the study carried out by Louw (2018) was of 15 frogs per $30m^2$. It was then difficult to know what the probability of inclusion was for these respective sites. Thus, in this study, a model-based approach was used to obtain an abundance estimate of *A. lightfooti*.

In a model-based approach, the method of selecting sampling locations is not a requirement as long as the element of randomness is introduced via the chosen model (Brus, 2010). The model-based approach relies on models that fit the data well. Subsequently, the covariates can then be used to predict density outside the sampled area. Thus, in this case, it does not matter which sites were sampled as long as the important environmental gradients are adequately sampled and a model that describes the variation in the density and occupancy of the frog can be found. If part of the environmental space is not sampled, there are then risks of extrapolation.

As it turns out, the null model was the most parsimonious at describing *A. lightfooti* density. The null model thus represents an estimate of mean density for the sampled sites (sample mean). Therefore, all variations found in the hurdle model predictions come from the variations in the predicted presence probabilities. The null model then just scales these probabilities by a constant. That constant being the mean number of frogs given they are present. I found that the predicted presence probabilities also tend towards the sample mean – with on average a 50% chance of encountering frogs at any random site on the Peninsula which in this case seems relatively high. In general, when a model is not good, the predictions tend towards the sample mean. In that case, the sampling design becomes critical as it heavily influences the sample mean.

The density survey was done using a stratified sampling design so that there were more samples at sites where the frogs were more likely to be present (see section 2.2). If the observed presence patterns followed the expected patterns, we would expect on average a higher probability of finding the species across sites (due to the overrepresented high probability sites in the data). One advantage of the model-based approach is that when the model has been shown to produce biases estimates, the model can be adjusted to give a more accurate estimation of the estimate of interest (Shi et al., 2016). In this case, in order to adjust for the imbalanced sampling – a weighted average would then reduce the estimated presence probabilities. In the case of *A. lightfooti*, the averaged observed presence probabilities were the same across strata. Thus, using a weighted average as adjustment to the estimates in each stratum would not bring any changes to the abundance estimate. The estimate obtained using the model-based approach should be considered with caution as it is very likely to be an overestimate of the true abundance.

Given the stratified sampling design used, I attempted to provide an unbiased design-based estimate of abundance of *A. lightfooti* which totalled to approximately 800 000 individuals across the Cape Peninsula. Using auxiliary

information available on the survey data, I was able to identify those sites that were not initially sampled, that is, sites that were located by walking in circle around the initially selected site (Appendix A). The density estimates from those sites were excluded from the calculations. Then using stratum-specific inclusion probabilities, a weighted mean of the densities across sites that were initially sampled was computed. However, this estimate should also be considered with caution as the standard errors associated with the density estimates cannot be accommodated in a design-based framework.

Furthermore, the frog density estimates were obtained by converting call density to frog density using a single estimate of call rate and the latter might not be constant throughout the season. Another source of uncertainty is the density estimates themselves which were computed from an auxiliary analysis accompanied with a degree of uncertainty. Modelling these data points without propagating the uncertainty could lead to deceptive estimates and overstate the precision of the reported abundance estimate. Although I accounted for this uncertainty during the modelling stage of the analysis (see section 4.4), one should carefully consider this global population.

From the resulting prediction of the combined model (Figure 6.1 – Left panel), regions of high density estimates are associated with the presence of wetland areas. These findings are expected, given that the species is vulnerable to dehydration and therefore require moist habitats (Channing, 2004). Results also revealed areas with the highest uncertainty in expected density estimates in the presence of large wetland areas (between 6220000N and 6225000N in Figure 5.1 Left panel). Moreover, spikes of high densities scattered among much lower densities could be also due to overpopulated habitats where species emigrate to a neighbouring site (in this case grid cell), which often are of lower quality (Gaston, 2003).

Large areas of moderate density estimates were observed in cooler montane regions. *A. lightfooti* are known to occur from sea level to 1000m (Channing, 2004) and lower pressure at higher altitudes causes the temperature to be colder on top of a mountain (Whiteman, 2000) which seems to be preferred by the species. Density estimates were generally lower on the southern part of the Cape Peninsula, where the MESS map (Figure 6.3 – left panel) indicates a dissimilar environmental space from the surveyed points onto which the model is predicting. Those areas of predictions should be carefully interrogated.

Moreover, the results demonstrated that incorporating a habitat suitability covariate derived from presence-only data from multiple sources did not seem to improve the estimation of density patterns of *A. lightfooti*. The leave-one-out cross-validation procedure resulted in higher cumulated mean squared error whenever the habitat suitability index map was incorporated as a covariate. This result is surprising since the presence-only derived habitat suitability should be informative on where the species occurs. So is the hurdle model perhaps describing a different process compared to the SDMs? Moreover, if the abundant-centre hypothesis is correct, then we would expect the habitat suitability map to be informative to both parts of the hurdle model, i.e. the part describing occurrence and the part describing density.

However, the results suggest that the habitat suitability map was uninformative, neither on the probability of occurrence nor on density where the species occurs. It seems that the locations of where the species are now are not well predicted by where they were previously found. Habitat dynamics due to natural perturbations are customary. The Cape Peninsula experienced its biggest fires on record in January 2000 and March 2015 (Slingsby, 2015). After fires, animals disappear or breed up from the small numbers that are still present. They can also recolonize in different places (for example, at different altitude) that can take a different amount of time (Hossack and Corn, 2007). Given the timeframe of the opportunistically collected data of *A. lightfooti*, it is not surprising to note that

the species were absent after the fires in areas where they were previously observed.

Moreover, an important factor in the reproduction of *A. lightfooti* is the presence of seepage areas in which they breed which are maintained by extended rainfall (Measey et al., 2017). Between 2015 and 2017, the South-western cape region of South Africa experienced three of its lowest rainfall years on record (Conservation, 2018). Thus, species such as *A. lightfooti* which rely on sites that offer moist conditions were likely to be affected by the drying climate. In this case, given that the density data were surveyed mostly during the drought period – suitable sites at which the species were previously found might not exist anymore.

Similar to the findings in Chapter 4, the best performing hurdle model suggests that the distribution and density of the Lightfoot's moss frog may be a function of different processes. In fact, despite fitting additional models to the density data in the aim of finding a better fit, the data always supported a null model suggesting that one is better off at just predicting constant density where the species occurs but at the cost of some bias.

Species data used for SDM frequently come from multiple sources (Araujo and Guisan, 2006; Austin, 2007). Most of these data consists of historical occurrence records (Tingley and Beissinger, 2009) as in the presence-only dataset used in this study. Another study where including lower quality data failed to improve models' performance was carried out by Reside et al. (2011) where they tested the performance of distribution models when incorporating historical presence-only data in addition to higher quality data to increase sample sizes.

In this study, a first-stage analysis of the presence-only data with no additional associated information, led to a habitat suitability index map. However, Pacifici et al. (2017) developed an approach where instead of using only higher quality dataset - they summarised information available with lower quality occurrence

data (such as information on sampling effort and number of observers present during sampling) and used them as covariates to improve predictions.

From this study, we learned that studies that rely on only opportunistically collected data should be considered with caution. Furthermore, the environmental covariates used in this study turned out to have limited predictive power, which resulted in a poor model. Given the need to rely on model-based inferences, more informative covariates would be needed to improve the reliability of the resulting abundance estimate. Moving forward, other important factors shaping density estimates should be considered such as site history and competition among species once occupancy is settled.

Chapter 7 - Conclusion

The advancement in technology and acoustic spatial capture-recapture methods have enabled the monitoring of visually cryptic but acoustically detectable species more closely and accurately than ever before. This study helped us gain insight into whether the two processes, that is, the distribution and relative abundance patterns of *A. lightfooti* are governed by the same environmental factors. Moreover, this study resulted in construction of the first Peninsula wide population density map of *A. lightfooti* while also investigating whether the opportunistically collected presence-only records can be used to improve and to produce the best possible population density map.

The hurdle model, by separately modelling the probability of presence and the densities given presence using different sets of explanatory variables have enabled us to gain insight whether the two processes were influenced by the covariates in different ways. Environmental factors influencing density may differ from those limiting distribution. Across all hypotheses, *A. lightfooti* density did not vary with measured environmental factors, with the null model being preferred across all cases. As expected, the occurrence of the species was mostly driven by the presence of wetlands. In contrast, predictions of density were only weakly related to these same environmental factors and in some cases contradicting one another. These findings suggest that the abundant-centre hypothesis might not be valid for all species.

In addition, I have been able to combine the results from the two analyses to estimate the expected density for *A. lightfooti* across its entire range. We now have the first quantitative population estimate of ~3.5 million calling male individuals across the Cape Peninsula. At the same time the study assessed the ability of using opportunistically collected presence-only records in combination with higher quality density data to improve the estimation of expected population-density surface of *A. lightfooti*. The presence-only records were

constructed into a habitat suitability map using an ensemble of species distribution models.

Seven different SDM algorithms were combined by taking the average prediction from each algorithm to build the ensemble model. Despite the overall good performance of all individual models, the techniques implemented have shown to vary in spatial predictions of the species' distribution. The ensemble model scored an AUC value higher than any individual SDMs fitted.

The findings suggested that ensemble forecast offers an improvement on any single model and its use for predictive species modelling is recommended. The habitat suitability map was then integrated in the modelling framework as a covariate in order to improve the estimation of expected population-density surface of *A. lightfooti* which resulted in being uninformative. From this study we learn that opportunistically collected occurrence records should be used cautiously when predicting species distributions.

References

1995. *STATISTICA for Windows.* Tulsa: StatSoft, Inc.

Abdi, H. 2010. Coefficient of variation. *Encyclopedia of Research Design*, 1), pp 169-171.

Abraham, A. 2005. *Artificial neural networks*, London: In handbook of measuring system design.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), pp 716-723.

Anderson, D. R., Burnham, K. P., White, G. C. & Otis, D. L. 1983. Density estimation of small-mammal populations using a trapping web and distance sampling methods. *Ecology*, 64(4), pp 674-680.

Anderson, D. R., Link, W. A., Johnson, D. H. & Burnham, K. P. 2001. Suggestions for presenting the results of data analyses. *Journal of Wildlife Management*, 65(3), pp 373-378.

Araujo, M. B. & Guisan, A. 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33(10), pp 1677-1688.

Araújo, M. B. & New, M. 2007. Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, 22(1), pp 42-47.

Araújo, M. B., Whittaker, R. J., Ladle, R. J. & Erhard, M. 2005. Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, 14(6), pp 529-538.

Austin, M. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, 200(1), pp 1-19.

Austin, M., Belbin, L., Meyers, J., Doherty, M. & Luoto, M. 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. *Ecological Modelling*, 199(2), pp 197-216.

Azzurro, E., Soto, S., Garofalo, G. & Maynou, F. 2013. Fistularia commersonii in the Mediterranean Sea: invasion history and distribution modeling based on presence-only records. *Biological Invasions*, 15(5), pp 977-990.

Balakrishnan, N. 1991. *Handbook of the logistic distribution*: CRC Press.

Baldanzi, S., McQuaid, C. D., Cannicci, S. & Porri, F. 2013. Environmental domains and range-limiting mechanisms: testing the Abundant Centre Hypothesis using Southern African sandhoppers. *PLoS One*, 8(1), pp e54598.

Barbet-Massin, M., Jiguet, F., Albert, C. H. & Thuiller, W. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, 3(2), pp 327-338.

Bascompte, J. 2009. Mutualistic networks. *Frontiers in Ecology and the Environment*, 7(8), pp 429-436.

Beebee, T. J. & Griffiths, R. A. 2005. The amphibian decline crisis: a watershed for conservation biology? *Biological Conservation*, 125(3), pp 271-285.

Bewick, V., Cheek, L. & Ball, J. 2005. Statistics review 14: Logistic regression. *Critical Care*, 9(1), pp 112.

Biernacki, C., Celeux, G. & Govaert, G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7), pp 719-725.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. & White, J.-S. S. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3), pp 127-135.

Borchers, D. 2012. A non-technical overview of spatially explicit capture–recapture models. *Journal of Ornithology*, 152(2), pp 435-444.

Borchers, D. L. & Efford, M. 2008. Spatially explicit maximum likelihood methods for capture–recapture studies. *Biometrics*, 64(2), pp 377-385.

Borchers, D. L. & Marques, T. A. 2017. From distance sampling to spatial capture–recapture. *AStA Advances in Statistical Analysis*, 101(4), pp 475-494.

Borchers, D. L., Stevenson, B., Kidney, D., Thomas, L. & Marques, T. A. 2015. A unifying model for capture–recapture and distance sampling surveys of wildlife populations. *Journal of the American Statistical Association*, 110(509), pp 195-204.

Breiman, L. 2001. Random forests. *Machine Learning*, 45(1), pp 5-32.

Breiman, L. 2017. *Classification and regression trees*: Routledge.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. 1984. Classification and regression trees. *Wadsworth, Belmont,CA*.

Brown, J. H. 1984. On the relationship between abundance and distribution of species. *The American Naturalist*, 124(2), pp 255-279.

Brus, D. 2010. *Design-based and model-based sampling strategies for soil monitoring* [Online]. Soil Science Centre, Wageningen University and Research Centre, P.O. Box47, 6700 AA Wageningen, The Netherlands. Available: https://library.wur.nl/WebQuery/wurpubs/fulltext/160490.

Buckland, S. & Elston, D. 1993. Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology*, 30(3), pp 478-495.

Buckland, S. T., Anderson, D. R., Burnham, K. P. & Laake, J. L. 2012. *Distance sampling: estimating abundance of biological populations*: Springer Science & Business media

Buntin, M. B. & Zaslavsky, A. M. 2004. Too much ado about two-part models and transformation?: Comparing methods of modeling Medicare expenditures. *Journal of Health Economics,* 23(3), pp 525-542.

Burnham, K. P. & Anderson, D. R. 2003. *Model selection and multimodel inference: a practical information-theoretic approach*: Springer Science & Business Media.

Burton, A. C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J. T., Bayne, E. & Boutin, S. 2015. Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. *Journal of Applied Ecology*, 52(3), pp 675-685.

Cáceres, M. D. & Legendre, P. 2009. Associations between species and groups of sites: indices and statistical inference. *Ecology,* 90(12), pp 3566-3574.

Calcagno, V. & de Mazancourt, C. 2010. glmulti: an R package for easy automated model selection with (generalized) linear models. *Journal of Statistical Software,* 34(12), pp 1-29.

Cameron, A. C. & Trivedi, P. K. 2013. *Regression analysis of count data*: Cambridge university press.

Capinha, C. & Anastácio, P. 2011. Assessing the environmental requirements of invaders using ensembles of distribution models. *Diversity and Distributions,* 17(1), pp 13-24.

Channing, A. 2001. *Amphibians of central and southern Africa*: Comstock Pub. Associates.

Channing, A. 2004. Genus Arthroleptella. Atlas and red data book of the frogs of South Africa, Lesotho and Swaziland. *Smithsonian Institute, Washington, DC, USA*, 206-219.

Chao, A. 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics,* 43(4), pp 783-791.

Chefaoui, R. M. & Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological modelling*, 210(4), pp 478-486.

Chen, X., Hu, B. & Yu, R. 2005. Spatial and temporal variation of phenological growing season and climate change impacts in temperate eastern China. *Global Change Biology*, 11(7), pp 1118-1130.

Clusella-Trullas, S. & Garcia, R. A. 2017. Impacts of invasive plants on animal diversity in South Africa: A synthesis. *Bothalia-African Biodiversity & Conservation*, 47(2), pp 1-12.

Collins, J. P., Crump, M. L. & Lovejoy III, T. E. 2009. *Extinction in our times: global amphibian decline*: Oxford University Press.

Conservation, T. 2018. *Global warming has already raised the risk of more severe droughts in Cape Town* [Online]. Available: https://theconversation.com/global-warming-has-already-raised-the-risk-of-more-severe-droughts-in-cape-town-107625.

Corn, P. S. 2005. Climate change and amphibians. *Animal biodiversity and Conservation*, 28(1), pp 59-67.

Cowling, R. M., Richardson, D. M. & Pierce, S. M. 2004. *Vegetation of southern Africa*: Cambridge University Press.

Cragg, J. G. 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, 39(5), pp 829-844.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. & Lawler, J. J. 2007. Random forests for classification in ecology. *Ecology*, 88(11), pp 2783-2792.

Davison, A. C. & Hinkley, D. V. 1997. *Bootstrap methods and their application*: Cambridge university press.

Dawood, A. & Channing, A. 2000. A molecular phylogeny of moss frogs from the Western Cape, South Africa, with a description of a new species. *Journal of Herpetology,* 34(3), pp 375-379.

Dawson, D. K. & Efford, M. G. 2009. Bird population density estimated from acoustic signals. *Journal of Applied Ecology,* 46(6), pp 1201-1209.

De'Ath, G. 2007. Boosted trees for ecological modeling and prediction. *Ecology,* 88(1), pp 243-251.

de Castro Godinho, M. B. & Da Silva, F. R. 2018. The influence of riverine barriers, climate, and topography on the biogeographic regionalization of Amazonian anurans. *Scientific reports,* 8(1), pp 3427.

DiCiccio, T. J. & Efron, B. 1996. Bootstrap confidence intervals. *Statistical Science,* 11(3), pp 189-212.

Dorazio, R. M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography,* 23(12), pp 1472-1484.

Du Preez, L. 2015. *A complete guide to the frogs of southern Africa*: Penguin Random House South Africa.

Efford, M. 2004. Density estimation in live-trapping studies. *Oikos,* 106(3), pp 598-610.

Efford, M. G., Borchers, D. L. & Byrom, A. E. 2009. Density estimation by spatially explicit capture–recapture: likelihood-based methods. *In:* Thomson D.L., C. E. G., Conroy M.J. (ed.) *Modeling demographic processes in marked populations.* Boston, MA: Springer.

Ehrlén, J. & Morris, W. F. 2015. Predicting changes in the distribution and abundance of species under environmental change. *Ecology Letters,* 18(3), pp 303-314.

Elith, J. 2015. *Predicting distributions of invasive species.*

Elith, J. & Graham, C. H. 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, 32(1), pp 66-77.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R. & Lehmann, A. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2), pp 129-151.

Elith, J., Kearney, M. & Phillips, S. 2010. The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 1(4), pp 330-342.

Elith, J. & Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40(677-697.

Elith, J., Leathwick, J. R. & Hastie, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), pp 802-813.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E. & Yates, C. J. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*, 17(1), pp 43-57.

Filz, K. J., Schmitt, T. & Engler, J. O. 2013. How fine is fine-scale? Questioning the use of fine-scale bioclimatic data in species distribution models used for forecasting abundance patterns in butterflies. *European Journal of Entomology*, 110(2), pp 311.

Fithian, W., Elith, J., Hastie, T. & Keith, D. A. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4), pp 424-438.

Fletcher, D., MacKenzie, D. & Villouta, E. 2005. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics*, 12(1), pp 45-54.

Franklin, J. 2010. *Mapping species distributions: spatial inference and prediction*: Cambridge University Press.

Gaston, K. J. 2003. *The structure and dynamics of geographic ranges*: Oxford University Press on Demand.

Gaston, K. J., Blackburn, T. M., Greenwood, J. J., Gregory, R. D., Quinn, R. M. & Lawton, J. H. 2000. Abundance–occupancy relationships. *Journal of Applied Ecology*, 37(39-59.

Gautheir, T. D. 2001. Detecting trends using spearman's rank correlation coefficient. *Environmental Forensics*, 2(4), pp 359-362.

Gazey, W. & Staley, M. 1986. Population estimation from mark-recapture experiments using a sequential bayes algorithm. *Ecology*, 67(4), pp 941-951.

Gerber, B. D., Karpanty, S. M. & Kelly, M. J. 2012. Evaluating the potential biases in carnivore capture–recapture studies associated with the use of lure and varying density estimation techniques using photographic-sampling data of the Malagasy civet. *Population Ecology*, 54(1), pp 43-54.

Gevrey, M., Dimopoulos, I. & Lek, S. 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling*, 160(3), pp 249-264.

Gillespie, D., Mellinger, D. K., Gordon, J., McLaren, D., Redmond, P., McHugh, R., Trinder, P., Deng, X. Y. & Thode, A. 2009. PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localization of cetaceans. *Journal of the Acoustical Society of America*, 125(4), pp 2547-2547.

Giraud, C., Calenge, C., Coron, C. & Julliard, R. 2016. Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics*, 72(2), pp 649-658.

Goh, A. T. 1995. Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, 9(3), pp 143-151.

Goodchild, M. F. 1993. Data models and data quality: problems and prospects. *Environmental Modeling with GIS*, 94-103.

Graunt, J. 1977. Natural and political observations mentioned in a following index, and made upon the bills of mortality. *Mathematical Demography.* Springer.

Grenouillet, G., Buisson, L., Casajus, N. & Lek, S. 2011. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography*, 34(1), pp 9-17.

Guisan, A., Edwards, T. C. & Hastie, T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 157(2), pp 89-100.

Guisan, A., Thuiller, W. & Zimmermann, N. E. 2017. *Habitat suitability and distribution models: with applications in R*: Cambridge University Press.

Guisan, A. & Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2), pp 147-186.

Hanley, J. A. & McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), pp 29-36.

Hanski, I. 1993. Three explanations of the positive relationship between distribution and abundance of species. *Species Diversity in Ecological Communities: Historical and Geographical Perspectives*, 108-116.

Harvey, P. H., Colwell, R. K., Silvertown, J. W. & May, R. M. 1983. Null models in ecology. *Annual Review of Ecology and Systematics*, 14(1), pp 189-211.

Hastie, T. & Tibshirani, R. 1990. *Generalized additive models*: Wiley Online Library.

Hefley, T. J. & Hooten, M. B. 2016. Hierarchical species distribution models. *Current Landscape Ecology Reports*, 1(2), pp 87-97.

Heilbron, D. C. 1994. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal,* 36(5), pp 531-547.

Heino, J. 2005. Positive relationship between regional distribution and local abundance in stream insects: a consequence of niche breadth or niche position? *Ecography,* 28(3), pp 345-354.

Herron, M. C. 1999. Postestimation uncertainty in limited dependent variable models. *Political Analysis,* 8(1), pp 83-98.

Hill, R. C. & Judge, G. 1987. Improved prediction in the presence of multicollinearity. *Journal of Econometrics,* 35(1), pp 83-100.

Hirschfeld, M., Blackburn, D. C., Doherty-Bone, T. M., Gonwouo, L. N., Ghose, S. & Rödel, M.-O. 2016. Dramatic declines of montane frogs in a central African biodiversity hotspot. *PloS One,* 11(5), pp e0155129.

Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X. 2013. *Applied logistic regression*: John Wiley & Sons.

Hossack, B. R. & Corn, P. S. 2007. Responses of pond-breeding amphibians to wildfire: short-term patterns in occupancy and colonization. *Ecological Applications,* 17(5), pp 1403-1410.

Howard, C., Stephens, P. A., Pearce-Higgins, J. W., Gregory, R. D. & Willis, S. G. 2014. Improving species distribution models: the value of data on abundance. *Methods in Ecology and Evolution,* 5(6), pp 506-513.

Hu, S. 2007. Akaike information criterion. *Center for Research in Scientific Computation,* 93(

Ishwaran, H. 2007. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics,* 1(2007), pp 519-537.

IUCN. 2018. The IUCN Red List of Threatened Species,

Jiménez-Valverde, A. 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, 21(4), pp 498-507.

Keeping, D. 2014. Rapid assessment of wildlife abundance: estimating animal density with track counts using body mass–day range scaling rules. *Animal Conservation*, 17(5), pp 486-497.

King, G. 1989. Event count models for international relations: Generalizations and applications. *International Studies Quarterly*, 33(2), pp 123-147.

Kipling, C. & Cren, E. 1984. Mark-recapture experiments on fish in Windermere, 1943–1982. *Journal of Fish Biology*, 24(4), pp 395-414.

Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1), pp 1-14.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D. & Morales, J. M. 2012. Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. *Ecology*, 93(11), pp 2336-2342.

Lek, S. & Guégan, J.-F. 1999. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, 120(2), pp 65-73.

Lettink, M. & Armstrong, D. P. 2003. An introduction to using mark-recapture analysis for monitoring threatened species. *Department of Conservation Technical Series A*, 28(5-32.

Li, X. & Wang, Y. 2013. Applying various algorithms for species distribution modelling. *Integrative Zoology*, 8(2), pp 124-135.

Liaw, A. & Wiener, M. 2002. Classification and regression by randomForest. *R news*, 2(3), pp 18-22.

Lobo, J. M., Jiménez-Valverde, A. & Real, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2), pp 145-151.

Louw, M. 2018. *Acoustic Spatial Capture-Recapture (aSCR) and the cryptic Cape Peninsula moss frog Arthroleptella lightfooti.* Master's Thesis, University of Stellenbosch, South Africa, University of Stellenbosch.

Loveridge, J. 1976. Strategies of water conservation in southern African frogs. *African Zoology,* 11(2), pp 319-333.

Lütolf, M., Kienast, F. & Guisan, A. 2006. The ghost of past species occurrence: improving species distribution models for presence-only data. *Journal of Applied Ecology,* 43(4), pp 802-815.

Maggini, R. 2011. *Species distribution models for conservation-oriented studies in Switzerland: filling data and tool gaps.* Doctoral dissertation, University of Lausanne, Switzerland.

Manel, S., Williams, H. C. & Ormerod, S. J. 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology,* 38(5), pp 921-931.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K. & Thuiller, W. 2009. Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions,* 15(1), pp 59-69.

Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D. & Tyack, P. L. 2013. Estimating animal population density using passive acoustics. *Biological Reviews,* 88(2), pp 287-309.

Martin, S. W., Marques, T. A., Thomas, L., Morrissey, R. P., Jarvis, S., DiMarzio, N., Moretti, D. & Mellinger, D. K. 2013. Estimating minke whale (Balaenoptera acutorostrata) boing sound density using passive acoustic sensors. *Marine Mammal Science,* 29(1), pp 142-158.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J. & Possingham, H. P. 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology letters,* 8(11), pp 1235-1246.

McCullagh, P. & Nelder, J. A. 1989. *Generalized linear models*: CRC press.

McCulloch, W. S. & Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), pp 115-133.

Measey, G. J., Stevenson, B. C., Scott, T., Altwegg, R. & Borchers, D. L. 2017. Counting chirps: acoustic monitoring of cryptic frogs. *Journal of Applied Ecology*, 54(3), pp 894-902.

Measey, G. J. & Tolley, K. A. 2011. Investigating the cause of the disjunct distribution of Amietophrynus pantherinus, the Endangered South African western leopard toad. *Conservation Genetics*, 12(1), pp 61-70.

Merow, C., Smith, M. J. & Silander, J. A. 2013a. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10), pp 1058-1069.

Merow, C., Smith, M. J. & Silander Jr, J. A. 2013b. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10), pp 1058-1069.

Mesgaran, M. B., Cousens, R. D. & Webber, B. L. 2014. Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity and Distributions*, 20(10), pp 1147-1159.

Miller, J. 2010. Species distribution modeling. *Geography Compass*, 4(6), pp 490-509.

Mills, L. S. 2012. *Conservation of wildlife populations: demography, genetics, and management*: John Wiley & Sons.

Minasny, B. & McBratney, A. B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & geosciences*, 32(9), pp 1378-1388.

Minter, L. 2004. *Atlas and red data book of the frogs of South Africa, Lesotho, and Swaziland*: Avian Demography Unit, University of Cape Town.

Mollet, P., Kery, M., Gardner, B., Pasinelli, G. & Royle, J. A. 2015. Estimating population size for capercaillie (Tetrao urogallus L.) with spatial capture-recapture models based on genotypes from one field sample. *PloS One*, 10(6), pp e0129020.

Mondol, S., Karanth, K. U., Kumar, N. S., Gopalaswamy, A. M., Andheria, A. & Ramakrishnan, U. 2009. Evaluation of non-invasive genetic sampling methods for estimating tiger population size. *Biological Conservation*, 142(10), pp 2350-2360.

Moore, D. L. & Vigilant, L. 2014. A population estimate of chimpanzees (Pan troglodytes schweinfurthii) in the Ugalla region using standard and spatially explicit genetic capture–recapture methods. *American Journal of Primatology*, 76(4), pp 335-346.

Mucina, L. & Rutherford, M. C. 2006. *The vegetation of South Africa, Lesotho and Swaziland*: South African National Biodiversity Institute.

Mullahy, J. 1986. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3), pp 341-365.

Naimi, B. & Araújo, M. B. 2016. sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography*, 39(4), pp 368-375.

Navas, C., Carvajalino-Fernández, J., Saboyá-Acosta, L., Rueda Solano, L. & Carvajalino Fernández, M. 2013. *The body temperature of active amphibians along a tropical elevation gradient: Patterns of mean and variance and inference from environmental data.*

Nelder, J. A. & Wedderburn, R. W. M. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), pp 370-384.

Nielsen, S. E., Johnson, C. J., Heard, D. C. & Boyce, M. S. 2005. Can models of presence-absence be used to scale abundance? Two case studies considering extremes in life history. *Ecography*, 28(2), pp 197-208.

Pacifici, K., Reich, B. J., Miller, D. A., Gardner, B., Stauffer, G., Singh, S., McKerrow, A. & Collazo, J. A. 2017. Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology,* 98(3), pp 840-850.

Pan, J.-X. & Fang, K.-T. 2002. Maximum likelihood estimation. *Growth Curve Models and Statistical Diagnostics.* Springer.

Péron, G. & Altwegg, R. 2015. The abundant centre syndrome and species distributions: insights from closely related species pairs in southern A frica. *Global Ecology and Biogeography*, 24(2), pp 215-225.

Peterson, A. T., Papeş, M. & Soberón, J. 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, 213(1), pp 63-72.

Peterson, A. T. & Soberón, J. 2012. Species distribution modeling and ecological niche modeling: getting the concepts right. *Natureza & Conservação,* 10(2), pp 102-107.

Phillips, S. J., Anderson, R. P. & Schapire, R. E. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3), pp 231-259.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. & Ferrier, S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1), pp 181-197.

Phillips, S. J., Dudík, M. & Schapire, R. E. A maximum entropy approach to species distribution modeling. Proceedings of the Twenty-First International Conference on Machine Learning, 2004. ACM, 83.

Pollock, K. H. 2000. Capture-recapture models. *Journal of the American Statistical Association*, 95(449), pp 293-296.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A. & McCarthy, M. A. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution,* 5(5), pp 397-406.

Potts, J. M. & Elith, J. 2006. Comparing species abundance models. *Ecological Modelling,* 199(2), pp 153-163.

R Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical computing.

R Development Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing

Raske, M., Lewbart, G. A., Dombrowski, D. S., Hale, P., Correa, M. & Christian, L. S. 2012. Body temperatures of selected amphibian and reptile species. *Journal of Zoo and Wildlife Medicine,* 43(3), pp 517-521.

Reside, A. E., Watson, I., VanDerWal, J. & Kutt, A. S. 2011. Incorporating low-resolution historic species location data decreases performance of distribution models. *Ecological Modelling,* 222(18), pp 3444-3448.

Ridgeway, G. 2007. Generalized Boosted Models: A guide to the gbm package. *Update,* 1(1), pp 2007.

Ridout, M., Demétrio, C. G. & Hinde, J. Models for count data with many zeros. Proceedings of the XIXth international biometric conference, 1998. International Biometric Society Invited Papers. Cape Town, South Africa, 179-192.

Riedmiller, M. & Braun, H. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. Neural Networks, 1993., IEEE International Conference on, 1993. IEEE, 586-591.

Rigby, R. & Stasinopoulos, D. 2009. A flexible regression approach using GAMLSS in R. *London Metropolitan University, London.*

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D. & Ripley, M. B. 2013. Package 'mass'. *CRAN Repos. Httpcran R-Proj. OrgwebpackagesMASSMASS Pdf.*

Ripley, B., Venables, W. & Ripley, M. B. 2016. Package 'nnet'. *R Package Version,* 7-3.

Roura-Pascual, N., Brotons, L., Peterson, A. T. & Thuiller, W. 2009. Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula. *Biological Invasions,* 11(4), pp 1017-1031.

Royle, J. A., Chandler, R. B., Sollmann, R. & Gardner, B. 2013. *Spatial capture-recapture*: Academic Press.

Royle, J. A., Magoun, A. J., Gardner, B., Valkenburg, P. & Lowell, R. E. 2011. Density estimation in a wolverine population using spatial capture–recapture models. *Journal of wildlife management,* 75(3), pp 604-611.

Saffari, S., Adnan, R. & Greene, W. 2012. Parameter estimation on hurdle Poisson regression model with censored data. *Jurnal Teknologi (Sciences and Engineering),* 57(SUPPL. 1), pp 189-198.

Sagarin, R. D. & Gaines, S. D. 2002. The 'abundant centre' distribution: to what extent is it a biogeographical rule? *Ecology Letters,* 5(1), pp 137-147.

Sagarin, R. D., Gaines, S. D. & Gaylord, B. 2006. Moving beyond assumptions to understand abundance distributions across the ranges of species. *Trends in Ecology & Evolution,* 21(9), pp 524-530.

Sawa, T. 1978. Information criteria for discriminating among alternative regression models. *Econometrica: Journal of the Econometric Society,* 1273-1291.

Scholz, F. 2014. Maximum likelihood estimation. *Wiley StatsRef: Statistics Reference Online.*

Seber, G. A. F. 1982. The estimation of animal abundance and related parameters.

Shi, Y., Cameron, C. J. & Heckathorn, D. D. 2016. Model-based and design-based inference: reducing bias due to differential recruitment in respondent-driven sampling. *Sociological Methods & Research*, 0049124116672682.

Slingsby, J. 2015. *Ecological impacts of fire on the Cape Peninsula* [Online]. SAEON. Available: http://www.saeon.ac.za/enewsletter/archives/2015/april2015/doc01.

Soberón, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, 10(12), pp 1115-1123.

Soberon, J. & Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, 2(1-10.

Soisalo, M. K. & Cavalcanti, S. M. 2006. Estimating the density of a jaguar population in the Brazilian Pantanal using camera-traps and capture–recapture sampling in combination with GPS radio-telemetry. *Biological conservation*, 129(4), pp 487-496.

Stefánsson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. *ICES journal of Marine Science*, 53(3), pp 577-588.

Stevenson, B. C., Borchers, D. L., Altwegg, R., Swift, R. J., Gillespie, D. M. & Measey, G. J. 2015. A general framework for animal density estimation from acoustic detections across a fixed microphone array. *Methods in Ecology and Evolution*, 6(1), pp 38-48.

Stockwell, D. 1992. *Machine learning and the problem of prediction and explanation in ecological modelling.* Doctoral dissertation, Australian National University.

Stohlgren, T. J., Ma, P., Kumar, S., Rocca, M., Morisette, J. T., Jarnevich, C. S. & Benson, N. 2010. Ensemble habitat mapping of invasive plant species. *Risk Analysis: An International Journal*, 30(2), pp 224-235.

Stokes, E. J., Johnson, A. & Rao, M. 2010. Monitoring wildlife populations for management. *Wildlife Conservation Society and the National University of Laos, Vientiane*.

Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44-47.

Symonds, M. R. & Moussalli, A. 2011. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, 65(1), pp 13-21.

Therneau, T. M., Atkinson, B. & Ripley, M. B. 2010. The rpart package. Chicago.

Thomas, L., Buckland, S. T., Burnham, K. P., Anderson, D. R., Laake, J. L., Borchers, D. L. & Strindberg, S. 2014. Distance sampling. *Wiley StatsRef: Statistics Reference Online*.

Thomas, L., Buckland, S. T., Rexstad, E. A., Laake, J. L., Strindberg, S., Hedley, S. L., Bishop, J. R., Marques, T. A. & Burnham, K. P. 2010. Distance software: design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology*, 47(1), pp 5-14.

Thompson, J. A., Bell, J. C. & Butler, C. A. 2001. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. *Geoderma*, 100(1-2), pp 67-89.

Thuiller, W. 2004. Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, 10(12), pp 2020-2027.

Thuiller, W., Georges, D., Engler, R., Breiner, F., Georges, M. D. & Thuiller, C. W. 2016. Package 'biomod2'.

Thuiller, W., Lafourcade, B. & Araujo, M. 2009. ModOperating manual for BIOMOD. *Thuiller W, Lafourcade B (2010) BIOMOD: species/climate modelling functions. R package version*, 1.1-5.

Tingley, M. W. & Beissinger, S. R. 2009. Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends in Ecology & Evolution*, 24(11), pp 625-633.

Turner, A. & Channing, A. 2017. Three new species of Arthroleptella Hewitt, 1926 (Anura: Pyxicephalidae) from the Cape Fold Mountains, South Africa. *African Journal of Herpetology*, 66(1), pp 53-78.

Turner, A. & De Villiers, A. 2007. Amphibians. *Western Cape Province State of Biodiversity. CapeNature Scientific Services, Cape Town*, 36-54.

Wackerly, D., Mendenhall, W. & Scheaffer, R. L. 2008. *Mathematical statistics with applications*: Cengage Learning.

Welsh, A. H., Cunningham, R. B., Donnelly, C. & Lindenmayer, D. B. 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, 88(1-3), pp 297-308.

White, G. C. 1982. *Capture-recapture and removal methods for sampling closed populations*: Los Alamos National Laboratory.

Whiteman, C. D. 2000. *Mountain meteorology: fundamentals and applications*: Oxford University Press.

Winkelmann, R. 2008. *Econometric analysis of count data*: Springer Science & Business Media.

Wood, S. N. 2006. *Generalized additive models: an introduction with R*: Chapman and Hall/CRC.

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H. & Veran, S. 2013. Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, 4(3), pp 236-243.

Yoccoz, N. G., Nichols, J. D. & Boulinier, T. 2001. Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution*, 16(8), pp 446-453.

Young, L. J. & Young, J. 2013. *Statistical ecology*: Springer Science & Business Media.

Zaniewski, A. E., Lehmann, A. & Overton, J. M. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, 157(2-3), pp 261-280.

Zurell, D., Elith, J. & Schröder, B. 2012. Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. *Diversity and Distributions*, 18(6), pp 628-634.

Appendix A

Table A.1 - Detailed List of Sources for Species occurrence data from 1933 to 2015.

	Sources	Number of occurrence records
Arthroleptella Lightfooti	The South African Institute for Aquatic Biodiversity (SAIAB)	10
	KwaZulu-Natal Museum	14
	Iziko South African Museum	26
	François Becker	254
	Cape Nature	112
	Alex Rebelo	49
	Animal Demography Unit (ADU)	27
	iNaturalist	2

Table *A.2* - Area under the receiver of characteristic curve during each cross-validation run across the three pseudo-absence datasets.

	Model	Run 1	Run 2	Run 3
PA 1	**GLM**	0.939	0.929	0.934
	GAM	0.937	0.926	0.9
	GBM	0.96	0.956	0.952
	RF	0.969	0.963	0.966
	CTA	0.9	0.92	0.909
	ANN	0.931	0.768	0.946
	MAXENT	0.918	0.872	0.85
PA 2	**GLM**	0.918	0.933	0.921
	GAM	0.925	0.928	0.905
	GBM	0.949	0.965	0.95
	RF	0.961	0.971	0.962
	CTA	0.904	0.926	0.872
	ANN	0.893	0.929	0.916
	MAXENT	0.926	0.916	0.927
PA 3	**GLM**	0.926	0.926	0.886
	GAM	0.895	0.917	0.906

	GBM	0.961	0.952	0.95
	RF	0.971	0.957	0.961
	CTA	0.914	0.928	0.933
	ANN	0.93	0.928	0.931
	MAXENT	0.931	0.929	0.925

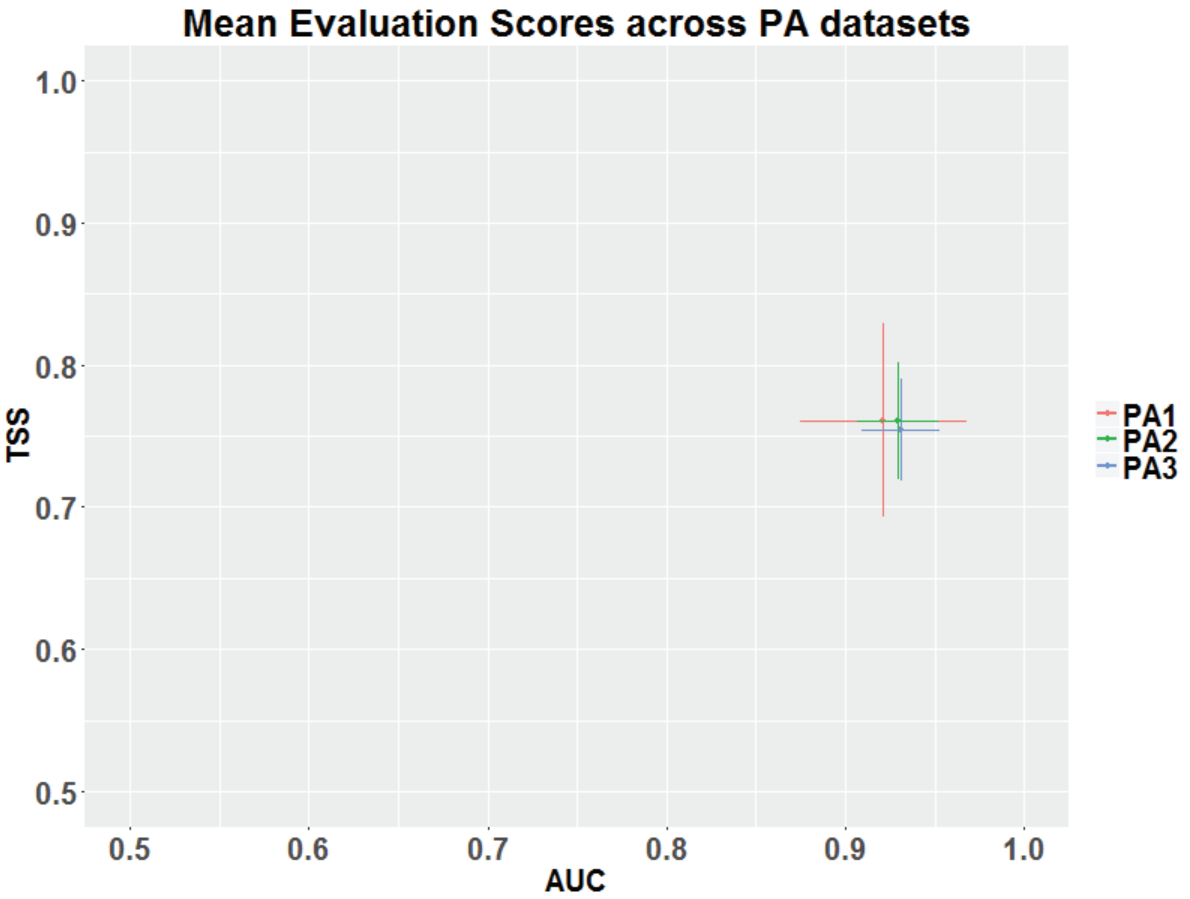


Figure A.3 - (a) Plot of the mean of the model evaluation scores across the different pseudo-absence datasets according to two different evaluation metrics, ROC (AUC) and TSS. The lines represent the associated standard deviations.

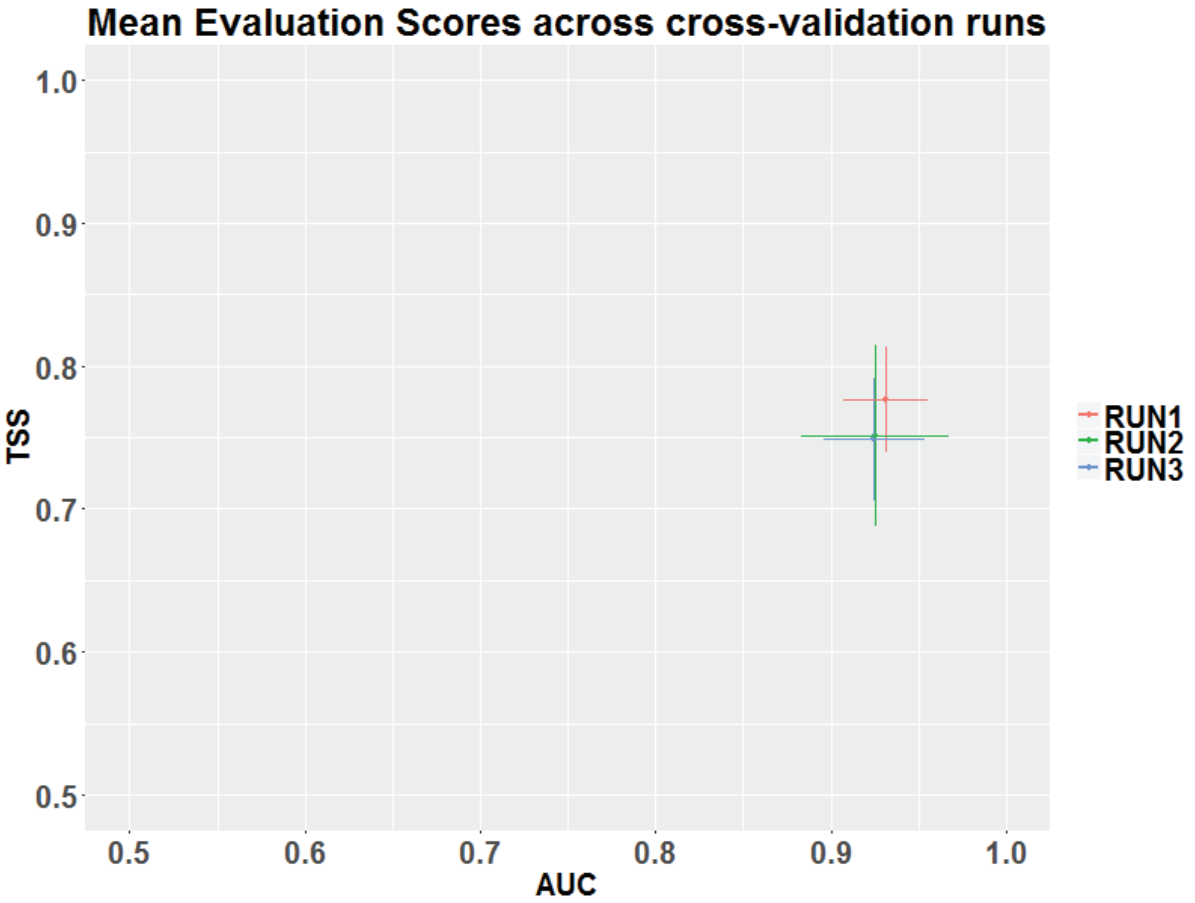


Figure A.3 - (b) Plot of the mean of the model evaluation scores across the different cross validation runs according to two different evaluation metrics, ROC (AUC) and TSS. The lines represent the associated standard deviations.

Table *A.4* - The average variable importance across the different pseudo-absences sets and cross-validation across the seven modelling algorithms.

	GLM	GAM	GBM	RF	CTA	ANN	MaxEnt
Elevation	0.00	0.12	0.01	0.13	0.00	0.29	0.09
Slope	0.04	0.05	0.03	0.04	0.12	0.00	0.12
Vegtypes	0.16	0.09	0.01	0.04	0.01	0.04	0.01
Vegsubtypes	0.13	0.05	0.01	0.04	0.00	0.03	0.05
Veg_communities_cowling	0.16	0.10	0.01	0.04	0.00	0.03	0.01
Northsouth	0.00	0.01	0.00	0.01	0.00	0.00	0.02
Jul_solar_radiation	0.00	0.04	0.01	0.02	0.00	0.11	0.07
Jan_solar_radiation	0.00	0.05	0.00	0.04	0.00	0.05	0.04
Eastwest	0.00	0.00	0.00	0.02	0.00	0.01	0.01
Roughness	0.00	0.06	0.00	0.03	0.00	0.05	0.02
Mean Max Summer Temperature	0.00	0.07	0.03	0.12	0.02	0.03	0.05
Mean Min Winter Temperature	0.44	0.30	0.79	0.34	0.71	0.23	0.33
TopographicPositionIndex30m	0.04	0.04	0.04	0.05	0.03	0.10	0.11
Wetlands	0.02	0.02	0.05	0.07	0.10	0.03	0.07

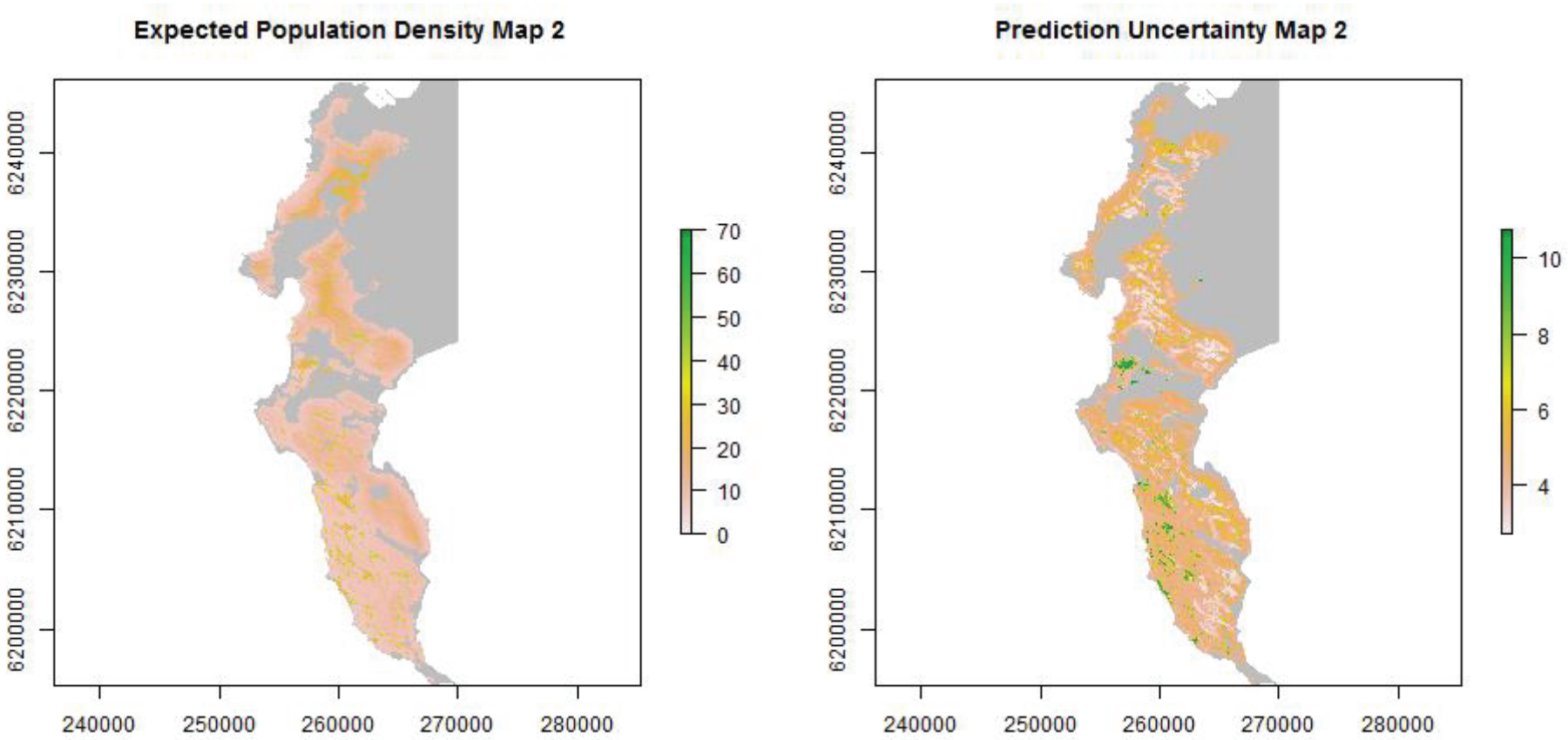


Figure A.5 - Expected population density estimates of the A. lightfooti across the Cape Peninsula. Green areas to low pink areas indicates high to lower estimates. In this map the habitat suitability map was used as a covariate in the logistic regression part of the Hurdle Model.

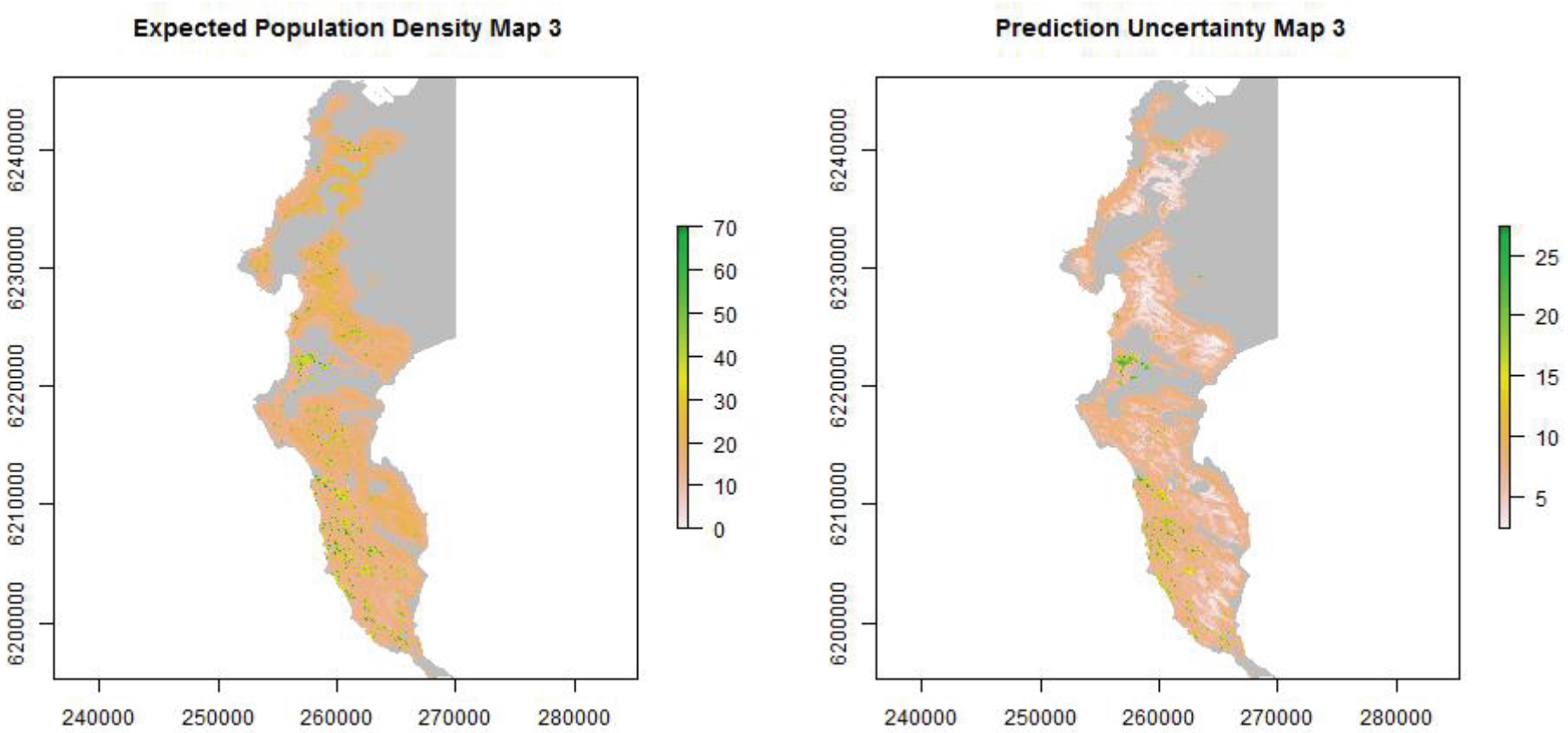


Figure A.6 - Expected population density estimates of the *A. lightfooti* across the Cape Peninsula. Green areas to low pink areas indicates high to lower estimates. In this map the habitat suitability map was used as a covariate in the gamma regression part of the Hurdle Model.

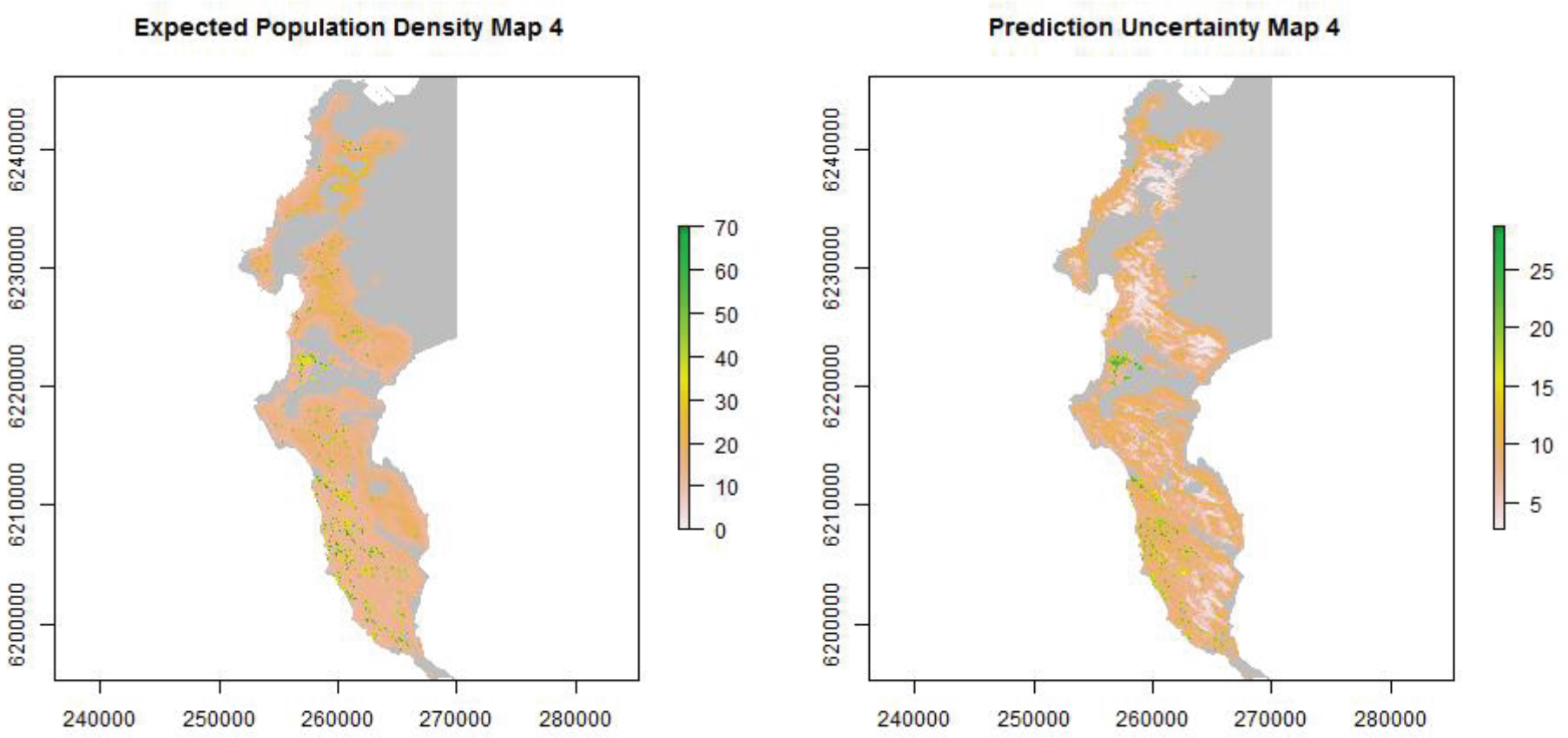


Figure A.7 - Expected population density estimates of the *A. lightfooti* across the Cape Peninsula. Green areas to low pink areas indicates high to lower estimates. In this map the habitat suitability map was used as a covariate in both part of the Hurdle Model.

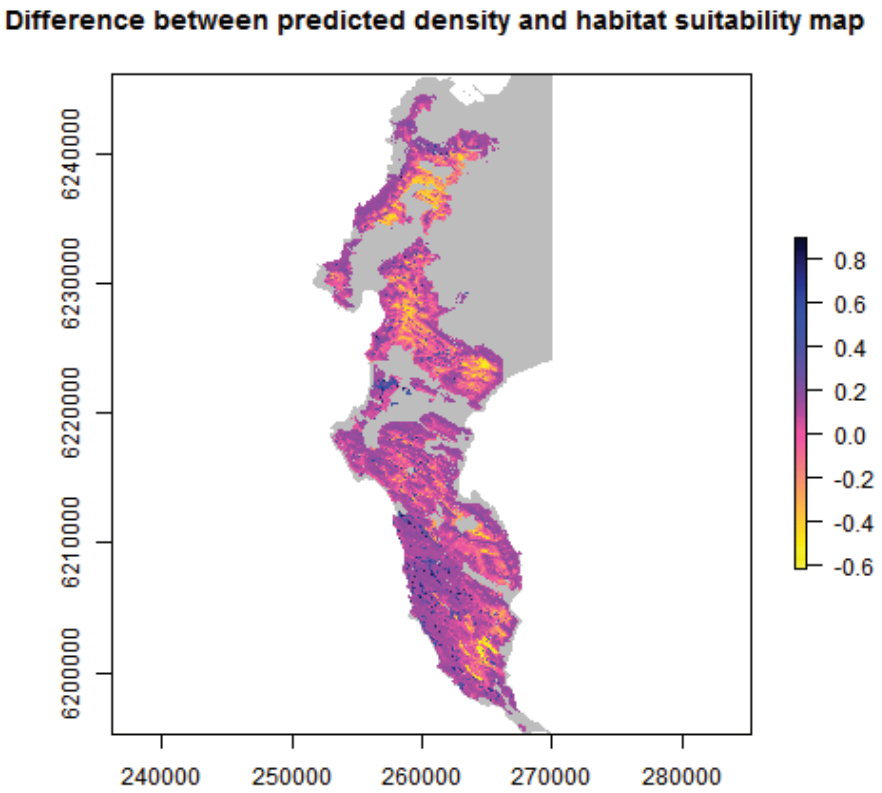


Figure A.8 - Mapped difference between the predicted presence probabilities from the presence/absence part of the hurdle model and the habitat suitability map. Positive values indicate areas where predictions from the hurdle model were higher that the ensemble SDMs predictions and vice versa.

Table A.9 Mean densities of initially selected sites across the four different stratums used during the survey.

Strata according to MaxEnt habitat suitability index	Mean density across each stratum (initially selected sites only)	Total number of cells in the whole study area falling into the stratum
(0.6 – 1)	9.111	5060
(0.4 – 0.6)	7.085	21496
(0.2 – 0.4)	2.947	72564
(0 – 0.2)	2.112	183996
Total		283116